Environmental Research Letters

LETTER

Machine learning assisted hybrid models can improve streamflow simulation in diverse catchments across the conterminous US

Goutam Konapala1,2, Shih-Chieh Kao1,2, Scott L Painter1 and Dan Lu1,3
1 Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, United States of America
2 Climate Change Science Institute, Oak Ridge National Laboratory, Oak Ridge, TN 37831, United States of America
3 Computational Sciences and Engineering Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, United States of America

E-mail: kaos@ornl.gov

Keywords: process-based hydrologic models, machine learning (ML), sacramento soil moisture accounting model (SAC), long short-term memory (LSTM) network

Supplementary material for this article is available online

Abstract

Incomplete representations of physical processes often lead to structural errors in process-based (PB) hydrologic models. Machine learning (ML) algorithms can reduce streamflow modeling errors but do not enforce physical consistency. As a result, ML algorithms may be unreliable if used to provide future hydroclimate projections where climates and land use patterns are outside the range of training data. Here we test hybrid models built by integrating PB model outputs with an ML algorithm known as long short-term memory (LSTM) network on their ability to simulate streamflow in 531 catchments representing diverse conditions across the Conterminous United States. Model performance of hybrid models as measured by Nash–Sutcliffe efficiency (NSE) improved relative to standalone PB and LSTM models. More importantly, hybrid models provide highest improvement in catchments where PB models fail completely (i.e. NSE < 0). However, all models performed poorly in catchments with extended low flow periods, suggesting need for additional research.

1. Introduction

Accurate simulation of streamflow and its sensitivity to various climate and anthropogenic factors has been a subject of research for many decades. To reasonably simulate streamflow, process-based (PB) hydrologic models have been the preferred choice (Hipsey et al 2015, Fatichi et al 2016). These models encode our understanding of catchment hydrologic processes through mathematical equations developed from various observations and experiments (Black 1996, Singh and Frevert 2005). As long as the key processes are captured and properly parametrized, PB models may provide reasonable streamflow simulation even outside the range of historical observations used for model calibration and validation (Klemes 1986, Carpenter and Georgakakos 2006, Seifert et al 2012, Abbott and Refsgaard 2012). However, formulation of PB models represents only a subset of the real processes in hydrologic cycle, leading to numerous constraints (Wagener and Gupta 2005, Vaché and Mcdonnell 2006, Renard et al 2010, Clark et al 2015). In cases where model bias exists due to insufficient scientific understanding or process representation, alternative approaches that can help remedy and enhance the streamflow simulation are of interest.

The rapid increase of hydrologic observations along with advances in computation capacity have led to powerful tools such as machine learning (ML; see Lecun et al 2015, Shen 2018 for review) that can make accurate simulation based on data alone. These ML tools have been applied in several fields of hydrologic like model parameterization (e.g. Gentine et al 2018, Krapu et al 2019), downscaling of hydrologic data (Abbaszadeh et al 2019), data retrieval from

© 2020 The Author(s). Published by IOP Publishing Ltd
remote sensing data (Karpatne et al. 2016, Ross et al. 2019, Cho et al. 2019), interpreting the hydrologic process (Jiang and Kumar 2019, Konapala and Mishra 2020), rainfall-runoff modeling (Kratzert et al. 2019, Xiang et al. 2020) with promising results. However, ML models often require longer training data sets to accurately capture the dynamics of complex processes in hydrologic cycle (e.g. Chen et al. 2018). Also, since these ML models are purely data-driven and have no process representation, they can lead to spurious and inaccurate predictions (e.g. Nayak et al. 2013, Lazer et al. 2014, Read et al. 2019), especially when predictions are outside of the range of data used for ML model training. This weakness strongly limits the applicability of ML algorithms in hydroclimate studies, in which both climate and land use patterns are projected to vary significantly and cannot be considered stationary.

PB and ML approaches are based on different philosophies but can complement each other with respect to their inherent strengths and limitations. The controlling hydrologic processes that preserve the mass and energy balance can be confidently captured by a PB model; while other secondary processes not incorporated in a PB model can be addressed by the data-driven ML algorithm (Shamseldin et al. 2007, Panda et al. 2010). Therefore, an appealing and effective strategy is to develop a hybrid combination of both approaches for enhanced hydrologic simulation. Following this general strategy, studies have integrated data-driven approaches such as the artificial neural networks (ANNs), support vector machines (SVMs) and genetic algorithms with hydrologic models (Chen and Adams 2006, Jain and Srinivasulu 2006, Corzo et al. 2009, Mekonnen et al. 2015, Humphrey et al. 2016, Young et al. 2017, Kumanlioglu and Fistikoglu 2019) and demonstrated better performance. Studies have also reported more accurate streamflow simulation by integrating PB outputs with ML methods such as the convolutional neural network and recurrent neural networks (RNN; Tian et al. 2018, Sun et al. 2019).

Even though these hybrid models have reported reasonable success in relatively smaller number of catchments, their applicability for a wide range of catchments with varying climate, land use and topographic conditions has not been reported. Testing their applicability over a wide variety of catchments would be beneficial for the following reasons. First, such large sample analysis provides better opportunities to improve hydrologic science (Ehret et al. 2014) by enabling better diagnosis of limitations, range of applicability and capabilities for extrapolation (Gupta et al. 2008, Martinez and Gupta 2011, Coron et al. 2012). Second, it enables rigorous statistical tests to quantify the robustness as well as the degrees of confidence of a proposed technique (Mathevet et al. 2006). Third, large samples would allow establishment of stronger relationships between model performance and catchment characteristics to provide insights into more detailed hydrologic behaviors (Sivapalan 2003, Götzinger and Bárdossy 2007, Oudin et al. 2010). To help better understand the strengths and limitations of hybrid models, efforts are needed to explore the performance of hybrid PB-ML models across different climate, land use and topographic conditions.

Among the ML approaches, long short-term memory (LSTM) network, which is a special type of RNN that can learn the long-term temporal dependencies between the variables (Hochreiter and Schmidhuber 1997), has been found more suitable for rainfall-runoff modeling (Kratzert et al. 2018). Kratzert et al. 2018 showed that a LSTM model can provide better daily streamflow modeling performance than the well-established Sacramento soil moisture accounting model (SAC; Burnash 1995) when coupled with the Snow-17 snow routine. Kratzert et al. 2019 further expanded the application of LSTM to more US watersheds to demonstrated LSTM's ability in simulating ungauged streamflow based on catchment characteristics. LSTM is also found to outperform other ML models such as SVM, ANN and RNN in terms of streamflow simulation across multiple time scales (Zhang et al. 2018, Kumar et al. 2019). In addition, Yang et al. 2019 demonstrated that a coupled LSTM-CalMa-Flood model can significantly enhance the performance of flood simulation. Given the success of LSTM in hydrologic applications, we further explore its applicability in a hybrid PB-ML modeling framework in this study. To build a hybrid PB-ML model, we use the SAC streamflow outputs from 531 US catchments in Catchment Attributes and Meteorological (CAMELS-US) dataset (Newman et al. 2014, Addor et al. 2017) as our large-sample hydrology dataset. We compare the performance of two different hybrid approaches and evaluate their performance against the standalone SAC and LSTM models. We then identify dominant catchment characteristics contributing to model performance and investigated functional relationship of dominant catchment attributes with the model performance using Random Forest algorithm’s dropout loss (Strobl et al. 2007, Grömping 2009) and accumulated local effect metrics (Apley et al., 2016) respectively.

2. Data and methodology

2.1. Data

Similar to Kratzert et al. 2018, we also select 531 CAMELS-US catchments which have no discrepancies in basin characteristics and have better model performance to train and compare different ML models. These CAMELS-US catchments have long-term streamflow observations, are minimally impacted by human activities, and cover a wide range of hydro-climatic conditions across the Conterminous US. For each catchment, we obtained 1980–2010 observed
daily precipitation, minimum and maximum daily temperature, solar radiation and vapor pressure as predictors for ML streamflow training and testing. These five meteorological predictors are spatially aggregated at each catchment from the 1 km resolution Daymet data set (Thornton et al. 2017). The streamflow observations are obtained from US Geological Survey National Water Information System. In addition to the observed streamflow time series, we also obtained simulated SAC streamflow from Newman et al. (2014) as the PB model output analyzed in this study. The SAC model was forced by the same catchment-averaged Daymet data for the period 1980–2010 and calibrated against 1980–1995 daily streamflow using the shuffled complex evolution global optimization routine. For further details on SAC model configuration and calibration, the readers are referred to Newman et al. (2014).

To understand the relationship between model performance and catchment features, we obtained characteristics related to: (1) climate, (2) topography and soil, (3) hydrology, (4) geology, and (5) vegetation for the same 531 catchments from Addor et al. (2017). These catchment characteristics were not used in the SAC model calibration, but provided as a complementary dataset to facilitate comprehensive multivariate catchment-scale assessments. The topographic characteristics (e.g. elevation and slope) were retrieved from Newman et al. (2014). Climatic indices (e.g. aridity and frequency of dry days) and hydrological signatures (e.g. mean annual discharge and basinflow index) were computed using observations provided by Newman et al. (2014). Additional soil (e.g. porosity and soil depth), vegetation (e.g. leaf area index and rooting depth), and geological (e.g. geologic class and subsurface porosity) characteristics were derived by Addor et al. (2017). The complete list of selected catchment characteristics, as well as details on the methods and data used to compute them, is provided in Supplementary Table S1. The catchments characteristics are summarized in Table S2. Further details on the CAMELS-US catchment characteristics are provided in Addor et al. (2017).

2.2. Long short-term memory network (LSTM)
A LSTM network is a special type of RNN that can store and regulate information over time and thus particularly suitable for simulating memory effects of meteorological inputs on streamflow (Shen 2018). This is important in streamflow simulation because some catchment processes such as snow accumulation, soil water storage and groundwater contribute significantly to streamflow at long time scales (Tague et al. 2008, Pellicciotti et al. 2010, Carey et al. 2013). The temporal memory effect is simulated by the combined use of cells and gates in LSTM (Hochreiter and Schmidhuber 1997), where the cells act as memory units and the gates control the information flow. Also, as opposed to the conventional RNN, LSTM does not have a problem with exploding and/or vanishing gradients, which allows it to efficiently map long-term dependencies between input and output sequences (Gers et al. 2000).

Given an input sequence with T time steps, \( x = [x[1],...,x[T]] \), where \( x[t] \) is a vector containing input features at any time step \( t \), the following equations describe the forward pass through the LSTM:

\[
\begin{align*}
    i[t] &= \sigma (W_i x[t] + U_i h[t-1] + b_i) \\
    f[t] &= \sigma (W_f x[t] + U_f h[t-1] + b_f) \\
    o[t] &= \sigma (W_o x[t] + U_o h[t-1] + b_o) \\
    g[t] &= \tanh (W_g x[t] + U_g h[t-1] + b_g) \\
    h[t] &= o[t] \circ \tanh (c[t]) \\
    c[t] &= f[t] \circ c[t-1] + i[t] \circ g[t]
\end{align*}
\]

where \( i[t], f[t] \) and \( o[t] \) are the input, forget, and output gates controlling the flow of memory, \( g[t] \) and \( h[t] \) are the cell update and recurrent inputs, \( c[t] \) represents the cell state, \( \{W_i, W_f, W_o, W_g\} \), \( \{U_i, U_f, U_o, U_g\} \) and \( \{b_i, b_f, b_o, b_g\} \) are the input, recurrent and biases weights, \( \sigma(\cdot) \) is the sigmoid function, \( \tanh(\cdot) \) is the hyperbolic tangent function, and \( \circ \) is element-wise multiplication. Based on the above formulation, the cell states \( \{c[t]\} \) characterize the memory of the system and can be modified by the forget gate \( f[t] \) to delete state memory, or by the input gate \( i[t] \) and cell update \( g[t] \) to add new information. In the latter case, the cell update is seen as the information that is added and the input gate controls which cells information is added. Finally, the output gate \( o[t] \) controls which information, stored in the cell states, is outputted. For a more detailed description, the readers are referred to Kratzert et al. (2018).

2.3. Construction of hybrid ML hydrologic models
In addition to the PB model output from Newman et al. (2014), we trained and tested three models to simulate daily streamflow. First, a stand-alone LSTM (Hybrid1) that simulates streamflow based on time series of the five Daymet variables (i.e. precipitation, maximum and minimum daily temperature, solar radiation and vapor pressure). Second, a hybrid LSTM (Hybrid2) that simulates
the PB model residual (i.e. simulated SAC streamflow minus observed streamflow) based on time series of the five Daymet variables. The estimated residual is then added to the PB model output to generate the corrected streamflow.

Since the SAC model is calibrated using daily data from 1980–1995, we use the same time period to train all three ML models to maximize the Nash–Sutcliffe efficiency (NSE; Nash and Sutcliffe 1970). NSE can be computed as:

\[
NSE = 1 - \frac{\sum_{t=1}^{T} (Q_{o,t} - Q_{e,t})^2}{\sum_{t=1}^{T} (Q_{o,t} - \bar{Q}_{o})^2}
\]  
(7)

where \(Q_{o,t}\) is mean of observed streamflow, \(Q_{e,t}\) is the modeled streamflow, and \(Q_{o,t}\) is the observed streamflow at time \(t\). NSE varies from \(-\infty\) to 1 with 1 representing the perfect match and negative values representing worse results than the mean value. We tested the model performance using 2000–2010 observed daily streamflow. For these three models (i.e. LSTM, Hybrid1 and Hybrid2) we used the same network architecture with 20 hidden states, 0.1 dropout rate, and 365 d as an input sequence with a single fully connected LSTM layer. The determination of the network hyperparameters have a significant influence on simulation results. We set the input sequence length of 365 d in order to capture the hydrological dynamics of a full annual cycle. We used a single layer network with relatively small hidden states plus dropout in order to avoid overfitting and enable the network to have a robust feature learning. We determined the specific values of hidden states and dropout through a number of test runs across different US catchments. The architecture used in this study proved to work well for these catchments and was therefore chosen to be applied here without further tuning.

### 2.4. Catchment features associated with model performance

To study the catchment features associated with the enhanced (or deteriorated) performance of the ML model with respect to the original PB model, we compute the change of NSE (e.g. \(\Delta NSE_{ML-PB} = NSE_{ML} - NSE_{PB}\)) during the validation period (2000–2010). We then perform a regression analysis of \(\Delta NSE\) using the catchment characteristics as predictors through the Random Forest algorithm (Breiman 2001). By using the Random Forest algorithm, we can rank the non-linear and non-parametric importance of each covariate to identify the dominant catchment characteristics associated with PB model performance using the dropout loss approach (Strobl et al 2007, Grömping 2009). Dropout loss is the decrease in random forest algorithm accuracy when a single covariate value is randomly shuffled. This procedure breaks the relationship between covariate and classification target, thus the decrease in accuracy score is indicative of how much the model depends on the tested predictor (i.e. higher dropout for more important variable). In general, an important covariate has a response function that displays interpretable features such as tipping points, well defined minima/maxima, or a slope (Murdoch, 2019). In contrast, variables with flat response functions (i.e. no change in NSE over the range of target variable) are relatively unimportant (Breiman 2001, Schwalm et al. 2017). Accumulated local effect (ALE; Apley 2016) plots based on the Random Forest algorithm can visualize this relationship between explanatory covariates and the probability of superior model performance. ALE gives the conditional effect of a covariate on the response variable and can get the isolated effect of each covariate even in the presence of correlation within covariates. Further details of dropout loss and ALE computations are provided in the supplementary information.

### 3. Results

#### 3.1. Model performance

Figure 1 describes various aspects of model performance as determined by NSE of the four tested models (PB, LSTM, Hybrid1 and Hybrid2). The spatial patterns of NSE across the four models are broadly similar. Weaker model performance is found across the Great Plains (in and around the states of South Dakota and Nebraska) and in the arid catchments in Texas and Arizona, where both LSTM and Hybrid1 models show comparatively higher NSE than PB and Hybrid2. On the other hand, all four models show stronger performance in the eastern and western US (figure 1(a)). The cumulative distribution functions (CDFs) of NSE values in figure 1(b) indicate that both hybrid models (Hybrid1 and Hybrid2) exhibit higher NSE values than the stand-alone PB and LSTM models, suggesting the overall enhanced performance through PB-ML hybrid modeling. Also, Hybrid1 is more correlated to LSTM (figure 1(c)), while Hybrid2 is more correlated to PB. The relatively higher correlation between LSTM and Hybrid1 (PB and Hybrid2) might be due to similar performance in the Great Plains and the arid catchments in Texas and Arizona.

We further explore the \(\Delta NSE\) of ML-assisted models (LSTM, Hybrid1 and Hybrid2) in figure 2. In figure 2(a), we classified all catchments by the original PB performance into four categories (\(NSE_{PB} \leq 0, 0 < NSE_{PB} < 0.5, 0.5 \leq NSE_{PB} < 0.75\) and \(NSE_{PB} > 0.75\)) and analyzed the spread of \(\Delta NSE_{LSTM-PB}, \Delta NSE_{Hybrid1-PB}, \Delta NSE_{Hybrid2-PB}\) through boxplots. Overall, the standalone LSTM performs better for catchments with poorer PB performance (\(NSE_{PB} \leq 0.5\)), but cannot outperform PB if the PB performance is already high (\(NSE_{PB} > 0.5\)). Both hybrid models can provide overall improvement but with different patterns. Hybrid1 that uses the PB simulated streamflow as the LSTM input behaves more similar to the standalone LSTM, while
Hybrid2 that uses the PB model residual as the LSTM input show more uniform but smaller enhancement across all catchments. It is worth mentioning that all ML-assisted models show highest improvement where the PB model has failed (i.e. catchments where $\text{NSE}_{\text{PB}} < 0$). However, as the PB model performance becomes better, even though there is a positive improvement in ML models, the magnitude of $\Delta\text{NSE}$ decreases. Among the ML models, Hybrid1 provides the highest median improvement over a wide variety of catchments across all four categories. Whereas, the standalone LSTM model has better performance than Hybrid2 when $\text{NSE}_{\text{PB}} \leq 0.5$, and vice versa.

The probability density functions (PDFs) of $\Delta\text{NSE}$ values are shown in figure 2(b), which indicate that the performance of Hybrid1 and Hybrid2 are overall higher than PB. It is interesting to see that there are fewer number of catchments with $\Delta\text{NSE} < 0$ in Hybrid2 (comparing to the higher number of catchments with $\Delta\text{NSE} < 0$ in LSTM and Hybrid1). This can also be observed in the spatial map in supplementary figure S1, which can be found online at [https://stacks.iop.org/ERL/15/104022/mmedia](https://stacks.iop.org/ERL/15/104022/mmedia), where Hybrid2 has a relatively smaller number of catchments with $\Delta\text{NSE} < 0$. This suggests the different effects of these two hybrid modeling approaches. While Hybrid1 may provide greater improvement, it may also lead to undesired deterioration in some catchments. Whereas, Hybrid2 can provide very consistent but weaker improvement. Based on the results, it seems that the most beneficial applications of PB-ML enhancement are for catchments with poorer PB.
performance \((\text{NSE}_{PB} \leq 0.5)\) through the Hybrid1 approach. In such cases, both standalone LSTM and Hybrid1 can largely improve the performance of streamflow simulation, suggesting the possibility of missing model processes captured by ML. Although both hybrid approaches can also improve streamflow simulation for catchments with better PB performance \((\text{NSE}_{PB} > 0.5)\), the improvement may be limited and could lead to undesired deterioration. These factors should be considered when evaluating the necessity to conduct streamflow modification through hybrid PB-ML modeling.

**3.2. Evaluation of catchment features associated with model performance**

Our next step is to understand what catchment features have the strongest influences on hybrid model performance. For this purpose, we assessed the changes in model performance of Hybrid1 with respect to the PB model \((\Delta \text{NSE}_{\text{Hybrid1-PB}})\) as well as to the standalone LSTM model \((\Delta \text{NSE}_{\text{Hybrid1-LSTM}})\). Investigating these two variables would shed light on where hybrid model might be advantage/disadvantage. We use dropout loss and conditional dependency to identify important catchment characteristics and further visualize their relationships with model performance. We also conducted similar analysis for Hybrid2 but found that the change of Hybrid2 performance do not correlate to catchment characteristics. Therefore, the results of Hybrid2 are not showed. Overall, Hybrid1 performs better than Hybrid2 in catchments with higher aridity, elevation and baseflow. Whereas, Hybrid2 performs better in catchments with higher precipitation, streamflow and high flow (supplementary figure S4).

Using dropout loss, in figure S2, we show that climate, soil and topography, and hydrologic characteristics have stronger effects on the change of model performance of Hybrid1 with respect to PB \((\Delta \text{NSE}_{\text{Hybrid1-PB}})\). In figure 3 we plot the accumulated local effect plots which visualize the contribution of the dominant catchment characteristics identified by dropout loss. As evident from figure S2, snow fraction has the highest dropout loss indicating that it is one most significant factor. Figure 3 highlights that with increasing snow fraction and elevation, the probability of better Hybrid1 performance increases. This suggests potential issues of PB model’s snow accumulation and ablation module (SNOW-17; Anderson 2006) for the applications on these CAMELS-US sites. In particular, we believe that the superior performance of Hybrid1 may be attributed to its ML component, which better captured the time lagged snow accumulation and melting process (Kratzert et al 2018).

We can also see that hybrid model improves the performance in regions of low precipitation and high aridity indicating their ability to capture the complex arid climate hydrology. Hybrid models have exhibited better performance in regions with high baseflow index. Factors that promote infiltration and recharge of subsurface storage will increase baseflow index, while factors associated with higher evapotranspiration will reduce baseflow index (Bloomfield et al 2009, Konapala and Mishra 2020). Therefore, in these catchments, the hybrid model was able to capture the slow groundwater drainage component which results in high baseflow index that might not be well represented in the PB model. It is interesting that ML models may distinguish these catchments based on data alone. This strength of LSTM was also reported by Kratzert et al (2019). It is also important to understand that the performance of hybrid model does decrease with low flow characteristics of duration and frequency. With the increase in low flow duration and frequency, there will be more irregular streamflow pulses which increases the complexity of streamflow series (Vogel and Fennessey 1994, Smakhtin 2001). Therefore, ML may not capture such complicated streamflow pattern. In addition, as the streamflow timing (i.e. half flow date) of the catchments shifts to end of water year the hybrid model performs better.

In figure 4, we show the functional relationship of top ten dominant catchment characteristics (figure S3) which have stronger effects on the change of model performance of Hybrid1 with respect to the standalone LSTM \((\Delta \text{NSE}_{\text{Hybrid1-LSTM}})\) through accumulated dependence plots. The results are quite different than figure 3, in which stronger effects are found for vegetation, climate, hydrology, and soil and topography features. As evident from figure S3, the 99th percentile root depth has the highest dropout loss, and figure 4 highlights that lower 99th percentile root depth would result in higher \(\Delta \text{NSE}_{\text{Hybrid1-LSTM}}\). Lower root depth indicates less infiltration and more water available for runoff and evaporation. Therefore, the hybrid model was able to capture this relationship better than the standalone LSTM model. Also, similar to figure 3, the hybrid model performs better than the LSTM model in arid, high elevation, low precipitation and snow dominant catchments. This suggests that the hybrid model may overcome the deficiencies in both PB and standalone LSTM models in case of these variables. However, unlike figure 3, we find that the hybrid model performs significantly better than the LSTM model in catchments with low magnitude of runoff ratio, streamflow and Q95. This suggests that the hybrid model may better represent the high streamflow generation mechanism that may not be well captured in either of the PB and standalone LSTM models. Finally, hybrid models perform better with respect to standalone LSTM models in catchments with less forest area fraction and vegetation.
Figure 3. The accumulated local plots which exhibit the functional relationship of the top six dominant catchment characteristics associated with changes in Hybrid1 to PB model performance ($\Delta$NSE$_{\text{Hybrid1-PB}}$). Variables are ordered by descending importance across the columns, each showing the averaged effect of a single dominant catchment characteristic on $\Delta$NSE$_{\text{Hybrid1-PB}}$ while holding the others constant.

Figure 4. Same as in figure 3, but for the top six dominant catchment characteristics associated with changes in Hybrid1 to LSTM model performance ($\Delta$NSE$_{\text{Hybrid1-LSTM}}$).

4. Discussion and conclusion

The superiority of LSTM over PB models in capturing hydrologic dynamics has been suggested in previous studies such as Kratzert et al (2018) and (Kratzert et al 2019). Our hybrid modeling approach that combines the output of a PB hydrological model with an LSTM network leads to further improvements of streamflow simulation. Specifically, the hybrid models significantly increased the median NSE relative to the standalone LSTM and PB models across diverse catchments in the US. On the important question of avoiding catastrophic failures (i.e. NSE < 0), the hybrid model performed similar to LSTM and significantly better than the PB model.

Our analysis using dropout loss and accumulated local effects indicated that the hybrid approach can overcome deficiencies in both PB and standalone ML models in capturing streamflow dynamics in snow-dominant, arid, and groundwater-dominant catchments. Specifically, the ML component in hybrid models is able to address the insufficient PB model process representation of snowmelt, streamflow generation in arid catchments, and groundwater contribution. Moreover, the PB component in the hybrid model may avoid the spurious simulation made by the standalone ML model by providing a first-order estimate based on the principles of water and energy balance at the catchment scale.

The spatial patterns of the hybrid model performance are in accordance with the ML model performance reported by Kratzert et al (2018). Even though the hybrid models provide significant improvements in NSE values, all models have deficiencies in the arid catchments in Great Plains, Texas and Arizona. Previous studies in those regions suggested that low PB model performance can be due to several reasons. First, these catchments have relatively sparse rain gauges leading to high forcing uncertainties. Second, since runoff only represents a small component within the overall water balance in arid catchments, small PB model errors may result in large relative differences (Newman et al 2017, Mizukami et al 2017). Finally, the PB models may miss some key processes.
related to surface–groundwater interactions, channel losses, and snowmelt (Hughes 2004, Rassam et al 2013, Kelleher et al 2015). Since the first two reasons deal with data availability and uncertainty, it is not surprising that the data-driven ML models can also be limited in these catchments. Overall, the NSE improvements by ML-assisted models (particularly Hybrid1) may be due to their abilities in capturing the key processes related to surface and ground water interactions which the PB model might have missed.

The hybrid models did have lower performance in the small fraction of catchments that had intermittent streamflow caused by frequent and extended low flow durations (figure 3). This may be due to a common challenge of both PB and ML models to capture the complex streamflow patterns. In catchments where neither the PB nor the standalone ML models produce satisfactory results, the hybrid model cannot ensure a better simulation.

These results confirm that hybrid models can combine the strengths of PB and ML models, with the PB models encoding scientific understanding of the hydrology system that is difficult to represent in the ML framework and the data-driven ML methods acting to add additional detail and remove biases created by incomplete or incorrect process description in the PB models. We envision this approach to be particularly valuable for the important class of applications involving projections of water availability in a changing climate condition, where standalone ML may have difficulty with out-of-sample prediction.

Although we found good success in coupling ML with a PB model, further advances in hybrid model algorithms can be expected. For instance, hybrid model training could use physics-based loss functions (Karatpatne et al 2016) that enforce water balance (Dingman 2015) or hydrological principles such as the Budyko hypothesis (Budyko 1974) to ensure that the model simulation is scientifically consistent with known streamflow generation mechanisms. Also, soil moisture and evapotranspiration observations, which have significant causal relationships with streamflow (Robinson et al 2008, Vereecken et al 2008, Penna et al 2011) but are difficult to use in PB models (Parajka et al 2006, Han et al 2012), could easily be introduced as additional input predictors. Our study used a common LSTM architecture across multiple catchments, but that could be extended to adopt different optimal architectures or hyper parameters for hydrologically distinct catchments. Alternative LSTM configurations could also be used, such as sequence-to-sequence modeling (Xiang et al 2020), which converts sequences of input forcings to sequences of output, or Bayesian LSTMs (Polson and Sokolov 2017), which characterizes the uncertainty of prediction.

Acknowledgments

This study was supported by the ExaSheds project within the US Department of Energy, Office of Science, Biological and Environmental Research Program. The research used resources of the Oak Ridge Leadership Computing Facility at ORNL, which is a DOE Office of Science User Facility. The authors are employees of UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy. Accordingly, the US government retains and the publisher, by accepting the article for publication, acknowledges that the US government retains a non-exclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US Government purposes.

Data availability statement

All data that support the findings of this study are included within the article (and any supplementary information files).

ORCID iD

Shih-Chieh Kao  https://orcid.org/0000-0002-3207-5328

References

Singh V P and Frevert D K ed 2005 Watershed Models (Boca Raton, FL: CRC Press)
Abbaszadeh P, Moradkhani H and Zhan X 2019 Downscaling SMAP radiometer soil moisture over the CONUS using an ensemble learning method Water Resour. Res. 55 324–44
Abbott M D and Refsgaard J C (eds) 2012 Distributed Hydrological Modelling vol 22 (Berlin: Springer)
Addor N, Newman A J, Mizukami N and Clark M P 2017 The CAMELS data set: catchment attributes and meteorology for large-sample studies Hydro. Earth Syst. Sci. 21 529–313
Anderson E 2006 Snow accumulation and ablation model—SNOW-17 National Weather Service Hydrology Laboratory, Silver Spring, MD (http://nws.noaa.gov/oh/hrl/nwshl/users_manual/part2/ pdf/22snow17.pdf)
Apley D W 2016 Visualizing the effects of predictor variables in black box supervised learning models arXiv Preprint, arXiv:1612.08464
Black P E 1996 Watershed Hydrology (Boca Raton, FL: CRC Press)
Bloomfield J P, Allen D J and Griffiths K J 2009 Examining geological controls on baseflow index (BFI) using regression analysis: an illustration from the Thames Basin, UK J. Hydrol. 373 164–76
Breiman L 2001 Random forests Mach. Learn. 45 5–32
Budyko M I 1974 Climate and Life (New York: Academic)
Burnash R J C 1985 The NWS river forecast system — catchment modeling Computer Models of Watershed Hydrology ed V P Singh (Lone Tree, CO: Water Resources Publication) pp 311–66
Carey S K, Boucher J L and Duarte C M 2013 Inferring groundwater contributions and pathways to streamflow
during snowmelt over multiple years in a discontinuous permafrost subarctic environment (Yukon, Canada). 

Hydrogeol. J. 21 67–77

Carpenter T M and Georgakakos K P 2006 Intercomparison of lumped versus distributed hydrologic model ensemble simulations on operational forecast scales. J. Hydrol. 329 174–85

Chen H, Engkvist O, Wang Y, olivecrona M and Blaschke T 2018 The rise of deep learning in drug discovery Drug Discovery Today 23 1241–50

Chen J and Adams B J 2006 Integration of artificial neural networks with conceptual models in rainfall–runoff modelling. J. Hydrol. 318 232–49

Cho E, Jacobs J M, Jia X and Kraatz S 2019 Identifying subsurface drainage using satellite big data and machine learning via google earth engine. Water Resour. Res. 55 8028–45

Clark M P et al 2015 A unified approach for process-based hydrologic modeling: 1. modeling concept Water Resour. Res. 51 2498–514

Coron L, Andréassian V, Perrin C, Lerat J, Vaze J, Bourqui M and Hendrick F 2012 Crash testing hydrological models in contrasted climate conditions: an experiment on 216 australian catchments Water Resour. Res. 48 W05552

Corzo G A, Solomatine D P, Hidayat M, de Wit M, Werner M, Uldenbroek S and Price R K 2009 Combining semi-distributed process-based and data-driven models in flow simulation: a case study of the meuse river basin Hydrol. Earth Syst. Sci. 13 1619–34

Dingman S L 2015 Physical Hydrology (Long Grove, IL: Waveland Press)

Ehret U et al 2014 Advancing catchment hydrology to deal with predictions under change Hydrol. Earth Syst. Sci. 18 649–71

Fatichi S et al 2016 An overview of current applications, challenges, and future trends in distributed process-based models in hydrology J. Hydrol. 537 45–60

Gentine P, Pritchard M, Rasp S, Reinaudi G and Yacalis G 2018 Could machine learning break the convection parameterization deadlock? Geophys. Res. Lett. 45 5742–51

Gers F A, Schmidhuber J A and Cummins F A 2000 Learning to forget: continual prediction with LSTM. Neural Comput. 12 2431–75

Grömping U 2009 Variable importance assessment in regression: linear regression versus random forest Am. Stat. 63 308–19

Gupta H V, Wagener T and Liu Y 2008 Reconciling theory with observations: elements of a diagnostic approach to model evaluation Hydrol. Process. 22 3802–13

Götzioner J and Bárdozy A 2007 Comparison of four regionalisation methods for a distributed hydrological model J. Hydrol. 333 174–84

Han E, Mervade V and Heathman G C 2012 Implementation of surface soil moisture data assimilation with watershed scale distributed hydrological model J. Hydrol. 416 98–117

Hipsey M R, Hamilton D P, Hanson D P, Carey C C, Coletti J Z, Read J S, Ibelings B W, Valesini F J and Brookes J D 2015 Predicting the resilience and recovery of aquatic systems: a framework for model evolution within environmental observatories Water Resour. Res. 51 7023–43

Hochreiter S and Schmidhuber J 1997 LSTM can solve hard long term lag problems Proc. of Conf. on Neural Information Processing Systems (Denver, CO) pp 473–9

Hughes D A 2004 Incorporating groundwater recharge and discharge functions into an existing monthly rainfall–runoff model Hydrol. Sci. J. 49 311

Humphrey G B, Gibbs M S, Dandy C G and Maier H R 2016 A hybrid approach to monthly streamflow forecasting: integrating hydrological model outputs into a bayesian artificial neural network J. Hydrol. 540 623–40

Jain A and Srinivasulu S 2006 Integrated approach to model decomposed flow hydrograph using artificial neural network and conceptual techniques J. Hydrol. 317 291–306

Jiang P and Kumar P 2019 Using information flow for whole system understanding from component dynamics Water Resour. Res. 55 8305–29

Karpate A et al 2016 Global monitoring of inland water dynamics: state-of-the-art, challenges, and opportunities Computational Sustainability ed J. Lässig, K. Kersting and K Morik (Cham: Springer) pp 121–47

Kelleher C, Wagener T and Mcglyn B V 2015 Model-based analysis of the influence of catchment properties on hydrologic partitioning across five mountain headwater subcatchments Water Resour. Res. 51 4109–36

Klemes V 1986 Operational testing of hydrological simulation models Hydrol. Sci. J. 31 13–24

Konapala G and Mishra A 2020 Quantifying climate and catchment control on hydrological drought in the continental United States Water Resour. Res. 56 e2018WR024620

Krapu C, Borsuk M and Kumar M 2019 Gradient-based inverse estimation for a rainfall–runoff model Water Resour. Res. 55 6625–39

Krkatrz F, Klotz D, Brenner C, Schulz K and Hrernnegger M 2018 Rainfall–runoff modelling using long short-term memory (LSTM) networks Hydrol. Earth Syst. Sci. 22 6005–22

Krkatrz F, Klotz D, Shalev Y, Kramauer G, Hochreiter S and Nearing G 2019 Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets Hydrol. Earth Syst. Sci. 23 5089–110

Kumanlioglu A and Fistikoglu O 2019 Performance enhancement of a conceptual hydrological model by integrating artificial intelligence J. Hydrol. Eng. 24 04019047

Kumar D, Singh A, Samui P and Jha R K 2019 Forecasting monthly precipitation using sequential modelling Hydro. Sci. J. 64 690–700

Lazen D, Kennedy R, King G and Vespiagnani A 2014 The parable of google flu: traps in big data analysis Science 343 1203–5

Lecun Y, Bengio Y and Hinton G 2015 Deep Learning Nature 521 436–44

Martinez G F and Gupta H V 2011 Hydrologic consistency as a basis for assessing complexity of monthly water balance models for the continental United States Water Resour. Res. 47 W12540

Mathevet T, Michel C, Andréassian V, Perrin C et al 2006 A bounded version of the nash-sutcliffe criterion for better model assessment on large sets of basins Large Sample Basin Experiments for Hydrological Model Parameterisation: Results of the Model Parameter Experiment – MOPEX V Andréassian, A Hall, N Chahinian and J Schaaeked (Wallingford: IAHS) pp 211–9

Mekonnen B A, Nazemi A, Mazurek K A, Elshobargy A and Putz G 2015 Hybrid modelling approach to prairie hydrology: fusing data-driven and process-based hydrological models Hydro. Sci. J. 60 1473–89

Mizukami N, Clark M P, Newman A J, Mizukami N, Clark M P, Newman A J, Wood A W, Gutmann E D, Nijssen B, Rakovec O and Samaniego L 2017 Towards seamless large-domain parameter estimation for hydrologic models Water Resour. Res. 53 8020–40

Murdoch W J, Singh C, Kumbier K, Abbasi-Asl R and Yu B 2019 Definitions of machine learning methods, and applications in interpretable machine learning Proc. Natl. Acad. Sci. USA 116 22071–80

Njeh J and Sutcliffe J V 1970 River flow river forecasting through conceptual models Part I – a discussion of principles J. Hydrol. 10 282–90

Nayak P C, Venkatesh B, Krishna B and Jain S K 2013 Rainfall–runoff modelling using conceptual, data driven, and wavelet based computing approach J. Hydrol. 493 57–67

Newman A J, Mizukami N, Clark M P, Wood A W, Nijssen B and Nearing G 2017 Benchmarking of a physically based hydrologic model J. Hydrometeorol. 18 2215–25

Newman A J, Sampson K, Clark M P, Bock A, Viger R J and Blodgett D 2014 A Large-Sample Watershed-Scale Hydrogeometrical Database for the Contiguous USA, UCA/ NCA, Boulder, CO
