Research papers

Simulating hydrologic pathway contributions in fluvial and karst settings: An evaluation of conceptual, physically-based, and deep learning modeling approaches

Admin Husic a,.*, Nabil Al-Aamery b, James F. Fox b

a Department of Civil, Environmental and Architectural Engineering, University of Kansas, United States
b Department of Civil Engineering, University of Kentucky, United States

ARTICLE INFO

Keywords:
Hydrologic modeling
Machine learning
LSTM
Rainfall-runoff
Process-based
Data-driven

ABSTRACT

Hydrologic models are robust tools for estimating key parameters in the management of water resources, including water inputs, storage, and pathway fluxes. The selection of process-based versus data-driven modeling structure is an important consideration, particularly as advancements in machine learning yield potential for improved model performance but at the cost of lacking physical analogues. Despite recent advancement, there exists an absence of cross-model comparison of the tradeoffs between process-based and data-driven model types in settings with varying hydrologic controls. In this study, we use physically-based (SWAT), conceptually-based (LUMP), and deep-learning (LSTM) models to simulate hydrologic pathway contributions for a fluvial watershed and a karst basin over a twenty-year period. We find that, while all models are satisfactory, the LSTM model outperformed both the SWAT and LUMP models in simulating total discharge and that the improved performance was more evident in the groundwater-dominated karst system than the surface-dominated fluvial stream. Further, the LSTM model was able to achieve this improved performance with only 10–25% of the observed time-series as training data. Regarding pathways, the LSTM model coupled with a recursive digital filter was able to successfully match the magnitude of process-based estimates of quick, intermediate, and slow flow contributions for both basins (\( p \) ranging from 0.58 to 0.71). However, the process-based models exhibited more realistic time-fractal scaling of hydrologic flow pathways compared to the LSTM model which, depending on project objectives, presents a potential drawback to the use of machine learning models for some hydrologic applications. This study demonstrates the utility and potential extraction of physical-analogues of LSTM modeling, which will be useful as deep learning approaches to hydrologic modeling become more prominent and modelers look for ways to infer physical information from data-driven predictions.

1. Introduction

Hydrologic modeling is a crucial tool for managing water resources and simulating water fluxes across – and within – the Earth’s surface (McMillan, 2021; Sit et al., 2020). Understanding the pathways by which rainfall inputs are routed to receiving waterbodies is a major goal for stakeholders and is crucial for managing water resources (Hammond et al., 2021). Numerous modeling approaches exist to tackle this problem, including process-driven and data-driven models (Bai et al., 2021; Kratzert et al., 2018). Model selection is influenced by existing system knowledge, availability of input data, parameter uncertainty, and access to computational resources (Shen, 2018; Sit et al., 2020). Questions remain regarding the tradeoffs of various modeling approaches for simulating hydrologic pathways and fluxes.

The prevailing paradigm of the last half-century of hydrologic modeling has been the development of “process-driven” conceptually and physically-based numerical models (Fatichi et al., 2016; Razavi, 2021). Process-driven models are constructed from empirical or analytical formulae that serve as proxies for physical phenomena, such as runoff, soil percolation, and groundwater recharge. However, developing process-based hydrologic models is exceptionally challenging as they are time-, data-, and computationally-intensive (Shen, 2018). In the last decade, rapid advances in “data-driven” machine learning models has provided a mechanism for equivalent or better
simulation of hydrologic systems with a fraction of the overall required resources (Razavi, 2021; Shen, 2018; Sit et al., 2020). In particular, the deep learning Long Short-Term Memory (LSTM) approach has proven especially fruitful for simulation of watershed dynamics because of its ability to ‘remember’ past system states (Kratzert et al., 2018). A common refrain regarding machine learning models is that their internal workings are difficult to discern and lack physical analogues (Razavi, 2021). To address this quite significant limitation, substantial work is being done to advance our interpretation of internal deep learning states (Kratzert et al., 2019a,b; Read et al., 2019). One particularly useful approach is to take existing, calibrated process-driven models and perform cross-comparison with machine learning outputs (Bai et al., 2021; Kratzert et al., 2018; Razavi, 2021). There exists the potential for developing approaches to benchmark data-driven estimates of hydrologic flow pathway contribution against process-driven outputs and overcome existing limitations of machine learning approaches.

Hydrologic pathways fall into three broad classifications: quick, intermediate, and slow flow (Husic et al., 2019b; Madsen et al., 2002). Each pathway represents a collection of spatially and temporally distinct fluxes that connect a water store to a catchment outlet (McMillan, 2022). Quick flow pathways are often turbulent and spatially discrete, respond rapidly to precipitation inputs, and account for low long-term storage of catchment water. Examples include surficial runoff, flow in tile-drains and underground caves, and point-source discharges. Slow flow is typically laminar and spatially diffuse, represents significant catchment storage, and sustains long-term recharge to streams and springs. Examples of slow pathways include deep aquifer recharge and flow through bedrock pores. Intermediate flow components bridge the spatial and temporal scales of quick and slow flow pathways and represent the dynamic interception and routing of flow to the catchment outlet or long-term storage. Examples of intermediate pathways include interflow and unsaturated flow in the vadose zone. Numerous methods exist for simulating the varying contribution of quick, intermediate, and slow flow to overall discharge, ranging from recursive digital filters (RDF) and recession curves for data-driven hydrograph separation (Singh and Stenger, 2018; Stoelzle et al., 2020) to conceptual or explicit numerical modeling (Al Aamery et al., 2021; Vansteenkiste et al., 2014). Few studies, however, have assessed the uncertainty associated with the choice of method for pathway modeling, which has implications for understanding the timing, extent, and connectivity of catchment fluxes.

Fig. 1. The location of the two basins of interest, South Elkhorn (fluvial-dominated) and Royal Spring (karst-dominated), situated near the city of Lexington in the Inner Bluegrass region of Kentucky.
The objective of this research was to assess the utility of deep learning models to simulate hydrologic flow path contributions in two neighboring systems, a surface-dominated fluvial stream and a subsurface-dominated karst spring. To compare outputs, we use the same input data to calibrate four different models: a conceptual model, a physically-based model, and two LSTM models (one each for the stream and spring). Thereafter, we tune and apply the models to a twenty-year flow period at both locations. We compare the quick, intermediate, and slow flow components generated from a modified LSTM model to those of the two process-driven numerical models. We perform statistical evaluation of the flow components, including auto-correlation, cross-correlation, and spectral analysis, to understand their temporal persistence and time fractal scaling. Lastly, we provide guidance on the applicability of deep learning models to simulate hydrologic flow path contributions in systems with varying surface–subsurface dominance over hydrology.

2. Materials and methods

2.1. Study site

To meet the objectives of this study, we selected two nearby basins in the Inner Bluegrass region of Kentucky, USA, with similar land use and topography but vastly different hydrologic pathway controls (Fig. 1). Royal Spring (58 km²) is a subsurface-dominated karst system whereas South Elkhorn (62 km²) is a surface-dominated fluvial system. The geology of the Inner Bluegrass region is comprised of phosphatic limestone of the Middle Ordovician period, which experiences chemical dissolution and erosion (Spangler, 1982). The variable dissolution of limestone bedrock results in the formation of karst features, such as sinkholes, caves, and conduits. In the Royal Spring basin, a series of karst sinkholes and swallets redirect surface runoff into a subsurface conduit system, approximately 20 m below the ground surface (Husic et al., 2017a,b), which routes flow to a perennial spring. For the most part, these features are absent in South Elkhorn (Mahoney et al., 2018b); thus, its hydrology is largely unaffected by karst. The climate in the region is temperate with a mean annual precipitation of 1,170 ± 200 mm and temperature of 13.0 ± 0.7 °C. Land use is primarily pasture/grassland with urbanized headwaters in both basins.

Royal Spring and South Elkhorn have been subject to extensive field assessment and numerical modeling over the last two decades to ascertain hydrologic controls in the basins. In Royal Spring, dye tracing (Paylor and Currens, 2004), electrical resistivity testing (Sawyer et al., 2015; Zhu et al., 2011), high frequency aquatic sensing (Husic et al., 2017a), water isotope tracing (Husic et al., 2019a), and numerical modeling (Husic et al., 2019b) have allowed us to develop a strong conceptual model of water storage and routing through the karst aquifer. Briefly, event precipitation is initially captured by the basin as distributed recharge through the soil or concentrated recharge through sinkholes, swallets, and depressions. Thereafter, distributed recharge percolates into the epikarst (the near-surface karst regolith), where it can be dynamically routed to the spring or deeper aquifer. During non-event periods, baseflow to the spring originates from both saturated and unsaturated zones of the aquifer. In South Elkhorn, numerical modeling of hydrology (Al Aamery et al., 2018, 2016; Ford and Fox, 2014) and stable isotope tracing (Ford et al., 2015; Mahoney et al., 2018a) have revealed the influence of storm events to runoff dynamics and soil storage to baseflow generation. Hydrologic delivery in South Elkhorn is largely unaffected by karst so the abstraction of surface runoff by karst holes is likely minimal. Thus, intense precipitation events generate considerable runoff that drains into the stream network. Thereafter, infiltrated water in the shallow and deep aquifers maintains elevated baseflow after event cessation.

2.2. Materials

To investigate hydrologic pathway controls, 20 years of daily discharge data were obtained from two United States Geological Survey gages: South Elkhorn (USGS 03289000) and Royal Spring (USGS 03288110). Meteorological input data for both basins, including precipitation and mean temperature, were obtained from the Bluegrass Airport (USW00093820). Potential evapotranspiration was estimated using the Penman-Monteith equation. Two existing calibrated numerical models were integrated into this study to serve as benchmarks for LSTM model performance (Fig. 2). These include a conceptually-based model for the Royal Spring basin (LUMP; Husic et al., 2019b) and a physically-based model for the South Elkhorn basin (SWAT; Al Aamery et al., 2018).

The LUMP model is a series of cascading linear reservoirs, each representing an entire water storage area for the basin, including the soil, epikarst, quickflow, and phreatic zones (Fig. 2a). Initially, precipitation input is separated into concentrated (X) and diffuse (1-X) fractions that contribute to quickflow and soil reservoirs, respectively. As these reservoirs overflow, they contribute to spring discharge ($Q_{spring}$) as quick flow ($Q_s$), epikarst flow ($Q_e$), and phreatic flow ($Q_p$). We infer the $Q_s$, $Q_e$, and $Q_p$ flow components to represent quick, intermediate, and slow recharge, respectively, to the subsurface-dominated spring. In its original study (Husic et al., 2019b), the LUMP model was calibrated to daily discharge from 10/2012 to 10/2014 (2 years) and validated from 10/2014 to 10/2016 (2 years). Model uncertainty was estimated from millions of Monte Carlo realizations covering the full range of potential parametrizations. In total, 1,176 model realizations were deemed successful in the original study and are used in this study. The simulation period in the original LUMP model paper covered 4 years, but has been extended to 20 years here, from 01/2000 to 01/2020. In summary, 2 years were used for calibration, 2 years for validation, and 16 years for testing.

The SWAT model identifies several Hydrologic Response Units (HRUs) – or parcels of the landscape with similar topography, pedology, and land use – and simulates a water balance for each HRU (Fig. 2b). The HRU output ($Q_{HRU}$) is a combination of surface runoff ($Q_s$), interflow ($Q_i$), and baseflow ($Q_b$) and contributes to a surface river network, which routes the flow to the basin outlet ($Q_{stream}$). We infer the $Q_s$, $Q_i$, and $Q_b$ flow components to represent quick, intermediate, and slow recharge, respectively, to the surface-dominated stream. In its original study (Al Aamery et al., 2018), the SWAT model was calibrated to daily discharge from 01/2006 to 01/2011 (5 years) and validated from 01/2011 to 01/2014 (3 years). Model uncertainty was assessed using the SWAT Calibration Uncertainty Program (SWAT-CUP SUFI-2). In total, 260 successful SWAT-CUP realizations are integrated from the original study. The simulation period in the original SWAT paper was 8 years but has been extended to 20 years, from 01/2000 to 01/2020. Thus, 5 years were used for calibration, 3 years for validation, and 12 years for testing.

2.3. Methods

2.3.1. Long-Short-Term memory model

An LSTM model is a subset of recurrent neural networks that have a dedicated memory cell state for long-term learning (Fig. 2c). Long-term learning is useful in hydrologic systems that have seasonal flow patterns, in particular snow hydrology, but may be useful in simulating the memory of groundwater-fed springs. We follow the outline for LSTM application to hydrology as laid out in Kratzert and others (2018). The LSTM model converts a network input sequence $x = [x_1, ..., x_T]$ with $T$ time steps, where each $x$ element is a vector containing the model inputs at a particular time step, $t$. The forward pass through the LSTM is described by the following set of equations:

$$ i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i), \quad (1) $$
\[ f[t] = \sigma(W_fx[t] + U_fh[t - 1] + b_f), \]  
\[ g[t] = \tanh(W_gx[t] + U_gh[t - 1] + b_g), \]  
\[ o[t] = \sigma(W_ox[t] + U_oh[t - 1] + b_o), \]  
\[ c[t] = f[t] \otimes c[t - 1] \oplus i[t] \otimes g[t], \]  
\[ h[t] = \tanh(c[t]) \otimes o[t]. \]  

where \( W, U, \) and \( b \) are learnable parameters for each gate, \( i[t], f[t], \) and \( o[t] \) are the input, forget, and output gates, respectively, \( g[t] \) is the cell input, and \( h[t - 1] \) is the recurrent hidden state input. \( c[t - 1] \) is the cell state from the previous step, \( \sigma(\cdot) \) is the sigmoid function, \( \tanh(\cdot) \) is the hyperbolic tangent function, \( \otimes \) is pointwise multiplication, and \( \oplus \) is pointwise addition. The outputs of the LSTM model are the \( h[t] \) and \( c[t] \) for the next step.

The basic flow of the LSTM network is that long-term memory cell states \( c[t] \) are modified by the forget gate \( f[t] \), which can delete certain states, and the input \( i[t] \) and cell update \( g[t] \) gates, which can add new information. Lastly, the output gate \( o[t] \) selects what information, stored in the cell states, is outputted. The output at the final time step \( h[T] \) is fed through a traditional dense layer to compute final discharge \( Q \), whether spring or stream, as:

\[ Q = W_0h[T] + b_0, \]  

where \( W_0 \) is the learnable weight matrix of the dense layer and \( b_0 \) is the bias term.

The learnable parameters \( W, U, \) and \( b \) are optimized by the LSTM model during the training period whereas the hyperparameters are defined prior to training. A generally agreed-upon method for configuring the hyperparameters presently does not exist in the literature. Our model was constructed within the MATLAB 2021 Deep Learning Toolbox and we use many of the default hyperparameter values with slight modifications derived from published literature to prevent overfitting (Bai et al., 2021). The customized hyperparameters are dropout rate = 0.1, numHiddenUnits = 32, MaxEpochs = 100, and miniBatchSize = 32.

The input variables for the two LSTM models – one for Royal Spring and one for South Elkhorn – are the same as those of the numerical models: daily precipitation, mean temperature, and potential evapotranspiration (Fig. 3). The outputs, for which the LSTM models were trained on, are the daily discharge from South Elkhorn \( Q_{SE} \) and Royal Spring \( Q_{RS} \). Zero-center normalization of training data was performed, providing for faster learning.

Process-driven models are typically split into calibration, validation, and testing splits (Shen et al., 2022). During calibration, model parameters are tuned such that the error between modeled and observed outputs is minimized. Thereafter, the model is validated to assess its performance during a period outside of its original calibration to evaluate the robustness of model fit. Model testing is the application of the model outside the original calibration and validation temporal space. In the machine learning community, the analogous terms for calibration,
validation, and testing are training, validation, and testing. Typically, over 70% of the dataset is used for training and validation with the remainder used for testing (Klotz et al., 2022; Kratzert et al., 2019a). In our case, the training periods for the Royal Spring and South Elkhorn LSTM models were selected to match the existing calibration periods of the LUMP and SWAT models, respectively, to provide like-for-like cross-model performance comparison. Likewise, the validation and testing periods of the LSTM models matched those of the LUMP and SWAT models. Thus, the LSTM models for South Elkhorn and Royal Spring were trained on 25% and 10% of the total dataset, respectively.

Model performance for all periods and all models was evaluated using the Modified Kling-Gupta Efficiency (KGE$^*$) (Clark et al., 2021).

The KGE$^*$ is calculated as:

$$\text{KGE}^* = 1 - \sqrt{\beta^2 + (\alpha - 1)^2 + (\rho - 1)^2},$$

where $\beta$, $\alpha$, and $\rho$ are the bias, standard deviation, correlation terms, respectively. $\beta$ is defined as $(\mu_s - \mu_o)^2/\sigma_o^2$, where $\mu_s$ is the simulated mean, $\mu_o$ is the observed mean, and $\sigma_o$ is the observed standard deviation. $\alpha$ is defined as $\sigma_s/\sigma_o$, where $\sigma_s$ is the simulated standard deviation. $\rho$ is the Pearson correlation between the observed and simulated flow. Larger positive values of KGE$^*$ represent better model performance with a KGE$^*$ of +1 indicating a perfect match between observed and simulated flow. Uncertainty in LSTM predictions was assessed by performing 1,000 realizations of the LSTM model—with dropout regularization enabled—for each basin and recording output predictions and performance metrics (Althoff et al., 2021). Thereafter, a prediction interval containing the inner 95th percentile was calculated from the 1,000 model realizations.

2.3.2. Hydrologic flow path analysis

In this study, we compare the hydrologic flow path estimates from process-driven (conceptual and physically based) models to those of
data-driven LSTM models. Broadly, we conceptualize spring and stream flow to be a combination of three components: quick, intermediate, and slow. The three flow components are simulated continuously by the LUMP model for Royal Spring and the SWAT model for South Elkhorn (Fig. 2). In the fluvial South Elkhorn, these flow components represent runoff, interflow, and baseflow, respectively. In Royal Spring, these components represent flow transported along conduit, fracture, and matrix karst features.

To estimate flow path contribution of the LSTM models, we apply an automated Lyne–Hollick (LH) recursive digital filter to modeled outputs. The LH filter applies signal processing techniques to a flow time series to separate high-frequency signals (inferred as quicker flow) from low-frequency signals (inferred as slower flow) (Ladson et al., 2013). Several passes of the LH filter can be performed with each subsequent pass separating the highest remaining frequency (Ladson et al., 2013). For each forward pass of the LH filter, the filtered high-frequency flow ($Q_1^l$) and low-frequency flow ($Q_1^s$) components can be calculated as:

$$Q_1^l = a(Q_{-1}^l) + (1 + a)(Q_0 - Q_{-1})/2,$$  

(9)

$$Q_1^s = Q_0 - Q_1^l,$$  

(10)

where $a$ is the dimensionless parameter (0 to 1), $Q_{-1}^l$ is the filtered low-frequency flow for the previous timestep, and $Q_0$ is the flow rate at the previous timestep. Larger values of $a$ correspond with smaller values of baseflow, and vice versa (Li et al., 2014). We use $a = 0.925$ for both Royal Spring and South Elkhorn, which is typically the default filter value (Li et al., 2014).

To separate the LSTM output signal into three flow components (i.e., quick, intermediate, and slow), we perform multiple LH filter passes. The first forward pass separates the bulk baseflow from the overall discharge (Eqs. (9)). The difference between the overall discharge and the first-pass baseflow yields the quick flow magnitude (Eqs. (10)). Subsequent passes further separate the baseflow into additional components, i.e., intermediate and slow flow. The number of passes needed to identify the lowest frequency component – in our case slow flow – can be calibrated to match results derived from alternate methods (Ladson et al., 2013), such as numerical modeling. For the Royal Spring and South Elkhorn systems, 3 and 8 forward-passes, respectively, were required to accurately distinguish the slow flow pathway. Thereafter, the intermediate flow was calculated as the difference between the 1st pass and the 3rd and 8th passes, respectively. Lastly, we investigated correlations between conceptual ($Q_{LUMP}$) and physically-based ($Q_{SWAT}$) and LSTM ($Q_{LSTM-RED}$) pathway contribution estimates using a $\rho$ value $>0.50$ to indicate strong correlation.

2.3.3. Statistical time series and frequency analyses

We used auto-correlation and cross-correlation analyses to better understand the temporal persistence of each flow pathway and pathway response to rainfall inputs (Bennett et al., 2013). The auto-correlation function indicates the memory effect of the system, and a value of 0.2 was used to represent the decorrelation lag time (Schuler et al., 2020). Cross-correlation can indicate the relationship between an uncorrelated cause, such as rainfall, and a subsequent effect, such as stream response (Bennett et al., 2013). We determine the response time of a flow pathway as the duration between rainfall inputs and the time-to-peak cross-correlation for the pathway.

The temporal structure of flow, such as the dominant frequencies and self-similarity of pathway contribution, were investigated with spectral analysis (Thompson and Katul, 2012). To transform the flow time series into frequency space, we calculated the power spectral density (PSD) using Welch’s fast Fourier transform algorithm. If a fractal process exists in the time-series, it will follow a power-law relationship between spectral power ($S$) and frequency ($f$) with temporal scaling of $S(f) \propto f^\beta$, where $\beta$ is a scaling exponent that represents signal persistence or memory of a signal (Jiang et al., 2020). This power-law relationship may appear as one or more linear segments in the log-log spectral density plot (Schuler et al., 2020). We pay particular attention to the ‘hydrologic regime’ portion of the PSD, which spans the daily-to-monthly return period (Thompson and Katul, 2012), and calculate $\beta$ associated with that period. We interpret the values of $\beta$ in the following way (Duran et al., 2020; Schuler et al., 2020):

- $1 < \beta < +1$: signal dominated by Gaussian noise, i.e., data pairs are independent from one another with $\beta = 0$ indicating random noise.
- $2 < \beta < -1$: signal characterized by anti-persistent Brownian noise, i.e., data pairs have poor correlation.
- $-3 < \beta < -2$: signal characterized by persistent Brownian noise, i.e., data pairs have high correlation and a memory effect.
- $\beta < -3$: signal is structured rather than stochastic, i.e., signal is strongly persistent.

We calculated $\beta$ for daily total, quick, intermediate, and slow flow for each SWAT, LUMP, and LSTM model realization using MATLAB (MathWorks, 2020).

3. Results

3.1. Comparative model performance evaluation

The LSTM models performed similarly – or better – compared to the benchmark models, SWAT and LUMP, for predicting stream and spring discharge, respectively, over a 20-year period in the Inner Bluegrass region of Kentucky, USA (Fig. 4). Regarding the fluvial South Elkhorn, the median KGE$^*$ scores for the SWAT calibration, validation, and testing periods were 0.71, 0.72, and 0.51, respectively. The corresponding median LSTM model scores were 0.65, 0.85, and 0.58, respectively, for the training, validation, and testing periods. Interestingly, the South Elkhorn LSTM model validation metrics exceeded those of training. One would expect the LSTM model performance to degrade outside of the training period. The reason this doesn’t occur is because the storm intensity and streamflow variability were much greater during model training ($Q_{\text{max}} = 73.25 \text{ m}^3/\text{s}$ and $\sigma = 3.37 \text{ m}^3/\text{s}$) than validation ($Q_{\text{max}} = 26.45 \text{ m}^3/\text{s}$ and $\sigma = 2.77 \text{ m}^3/\text{s}$), allowing for the model to capture the mean behavior more easily during validation. The out-sized impact of a few extreme events to skew hydrologic model performance are commonly reported (Knoben et al., 2019).

Regarding the karst Royal Spring, the median KGE$^*$ scores for the LUMP model were 0.66, 0.61, and 0.63 for the calibration, validation, and testing periods, respectively. Correspondingly, the Royal Spring median LSTM model scores were 0.74, 0.73, and 0.71 for the training, validation, and testing periods, respectively. While we constrained our training, validation, and testing periods to match the respective periods from the original LUMP and SWAT publications (Al Aamery et al., 2018; Husic et al., 2019b), there is potential for a more complete assessment of model evaluation, such as through a differential split-sample test to vary evaluation periods to representatively cover hot, dry, cold, and wet year variability in the dataset (Bai et al., 2021; Dakhlaoui et al., 2020). This point notwithstanding, all model structures (SWAT, LUMP, and LSTM) nonetheless exhibited ‘good’ performance (KGE$^* > 0.5$) during all evaluation periods.

The South Elkhorn and Royal Spring LSTM models performed particularly well considering that they were trained on only 25 % and 10 % of the study period, respectively. These training durations were selected to match the original calibration periods of the hydrologic models. Typically, over 70 % of the evaluation period is chosen for training a hydrologic dataset (Kratzert et al., 2018). Even with the smaller training period, the LSTM models do not appear to suffer from over-parameterization as their performance during the validation and training periods is similar to – or exceeds – that of the training period for both basins. To test how well the LSTM model for each basin would do
under traditional training splits (80% training and 20% testing), we conducted additional test runs. For South Elkhorn, the training and testing KGE values for the runs with more training data were 0.65 and 0.59, respectively, and for Royal Spring, they were 0.69 and 0.57, respectively. These values are similar, but not ‘better’ than the shorter training/validation/testing splits discussed earlier. Ostensibly, the KGE values slightly worsen as more data are integrated into model training, but this is primarily a result of how the KGE metric is constructed where bias ($\beta$), standard deviation ($\alpha$), and correlation ($\rho$) terms in Equation (8) may differ depending on the window of values considered in training, validation, and testing. Some caution should be taken in directly comparing KGE statistics when the duration and starting/ending points of evaluation periods differ. Nonetheless, our results suggest that – to match the performance of a conceptually or physically based hydrologic model – LSTM models may be able to use shorter training periods than typically suggested.

3.2. Modeling hydrologic pathway contribution

The hydrologic pathway that contributes most to streamflow in South Elkhorn – as indicated by SWAT modeling – is the intermediate flow component (Fig. 5), which constitutes 64.5 $\pm$ 4.5 % of total streamflow during the 20-year period. Intermediate flow for South Elkhorn is most analogous to shallow lateral recharge from hillslopes to stream networks. The next most prominent pathway for South Elkhorn is quick flow, which makes up 25.3 $\pm$ 3.5 % of total streamflow and represents surficial runoff to streams. The least prominent pathway for the surface system is slow flow, which represents deep aquifer recharge to the stream and accounts for 10.2 $\pm$ 2.6 % of total streamflow.

In Royal Spring, LUMP modeling results show a differing importance of flow pathways for the karst-dominated basin (Fig. 5). Intermediate flow for South Elkhorn is most analogous to shallow lateral recharge from hillslopes to stream networks. The next most prominent pathway for South Elkhorn is quick flow, which makes up 25.3 $\pm$ 3.5 % of total streamflow and represents surficial runoff to streams. The least prominent pathway for the surface system is slow flow, which represents deep aquifer recharge to the stream and accounts for 10.2 $\pm$ 2.6 % of total streamflow.

In Royal Spring, LUMP modeling results show a differing importance of flow pathways for the karst-dominated basin (Fig. 5). Intermediate flow is of second-importance and contributes 33.5 $\pm$ 21.5 % of total spring discharge and represents the drainage of water through fractures in surficial epikarst. Slow flow, representing deep recharge from matrix water storage, is the dominant flow pathway in Royal Spring,
representing 46.0 ± 20.6 % of spring flow. Lastly, quick flow – or flow routed through sinkholes, swallets, and conduits – is the pathway of least importance, accounting for 20.5 ± 8.4 % of spring discharge.

It is important to note that while a pathway may deliver relatively less water when integrated over the 20-year period, it can be temporally important during individual days, weeks, or seasons depending on spatiotemporal variability in recharge, soil moisture, and aquifer saturation. For example, in the South Elkhorn basin, runoff represents only 25.3 ± 3.5 % of total stream discharge, but on the 50 wettest days of the 20-year period, its contribution doubles to 53.7 ± 13.9 %. Likewise, in the Royal Spring basin, quickflow pathways drained only 20.5 ± 8.4 % of flow over the entire period, but that number increased to 42.3 ± 23.1 % for the 50 wettest days. Indeed, strong seasonal contributions are observed in both South Elkhorn and Royal Spring despite their stark differences in hydrology. Both basins show slow flow contributions that grow during the wet season (winter and spring) diminish during the dry season (summer and fall). On the other hand, intermediate flow connectivity remains high in both basins and is responsive to rainfall throughout the year and becomes relatively much more important in the dry season as deeper pathways are disconnected (see Apr-18 to Oct-18 in “South Elkhorn” – SWAT in Fig. 5).

The LSTM model predictions coupled to recursive digital filtering (QLSTM + RDF) successfully simulated the hydrologic pathway contribution predicted by conceptually (QLUMP) and physically-based (QSWAT) numerical models (Fig. 5). For both basins, only a single LH filter forward-pass was required to accurately extract the quick and bulk baseflow fractions. To further separate the bulk baseflow into intermediate and slow flow components, two additional forward passes were required for Royal Spring and seven additional forward passes were required for South Elkhorn. South Elkhorn requires more LH filter passes because its deeper, slower baseflow is entirely disconnected from the stream for large parts of the year (see “South Elkhorn – SWAT” in Fig. 5). Recursive digital filters struggle representing flow components that vanish for large parts of the year and accurate characterization requires multiple additional passes. Royal Spring, on the other hand, has some slow flow contribution even during periods of overall low flow.
magnitudes (see summer before Oct-18 in “Royal Spring – LUMP” in Fig. 5). Further, the slow flow contribution in the groundwater-fed spring (46.0 ± 20.6 %) is more prominent than in the surface-driven stream (10.2 ± 2.6 %), thus the imprint of its signal is more prominent and easier for recursive digital filters to identify.

In the surface-dominated South Elkhorn, LSTM + RDF flow contributions were strongly correlate to SWAT-predicted contributions for quick (\( \rho = 0.59 \)), intermediate (\( \rho = 0.61 \)), and slow (\( \rho = 0.61 \)) pathways. Likewise, in Royal Spring, LSTM + RDF predictions were even more correlated with LUMP-predicted quick (\( \rho = 0.55 \)), intermediate (\( \rho = 0.70 \)), and slow (\( \rho = 0.71 \)) contributions. Despite the relatively strong performance of the LSTM + RDF model for predicting flow path contribution, there were some notable limitations and sources of error in the method. In comparing \( Q_{\text{SWAT}} \) vs \( Q_{\text{LSTM + RDF}} \), it can be observed that, on some days, the LSTM + RDF model simulates > 5 mm d\(^{-1}\) of quick flow contribution when the SWAT model predicts no runoff (0 mm d\(^{-1}\)). This may occur because surface runoff (quick flow) in South Elkhorn usually occurs over just a single day, whereas the LSTM + RDF model artificially lengthens runoff duration due to how the LH filter is formulated. To further understand how the various models represent the temporal structure of the flow time series, we investigate fractal scaling and lag-time correlation in the next section.

### 3.3. Temporal structure and response of hydrologic pathways

The observed lag-time for the flow time-series to become decorrelated was 8 days in South Elkhorn compared to 64 days in Royal Spring (Fig. 6 and Table 1). These results were unsurprising as South Elkhorn is a flashy, runoff-dominated stream whereas Royal Spring is a groundwater-fed system that is primarily driven by baseflow. Regarding the model simulations, the SWAT model (4 days) indicated shorter decorrelation than observed in South Elkhorn, while the LSTM model (8 days) matched observations. In Royal Spring, both the modeled LUMP (68) and LSTM (57) decorrelation generally agreed with observations.

With an understanding of the auto-correlation of the overall flow time series, we next investigated the decorrelation times for the individual flow components (Fig. 6 and Table 1). For quick, intermediate, and slow flow components in South Elkhorn, the LSTM model (3, 19, and 88 days, respectively) predicted longer median temporal persistence of flow pathways than the SWAT model (2, 6, and 75 days, respectively). In Royal Spring, the behavior was more varied as the LSTM model (10, 16, and 79 days, respectively) predicted longer persistence of quick flow, but much shorter decorrelation timing of intermediate flow compared to the LUMP model (4, 56, and 77 days, respectively).

Power spectral scaling was present in the observed and modeled flow time series for both South Elkhorn and Royal Spring (Fig. 7 and Table 2). The scaling parameter (\( \beta \)) for the rainfall input was −0.20 (not plotted), indicating a generally noisy and unstructured signal. However, as this rainfall input is routed through physical watershed stores and discharged to basin outlets, it begins to follow more organized scaling: \( \beta = -0.92 \) for South Elkhorn observations and \( \beta = -1.45 \) for Royal Spring observations. The Royal Spring time series is likely more structured because the karst system attenuates the noisy rainfall input to a greater degree than the flashy South Elkhorn system. Regarding the modeling outputs, the SWAT model (\( \beta = -0.63 \)) indicated less structure than observed in South Elkhorn, while the LSTM model (\( \beta = -1.22 \)) suggested greater structure. In Royal Spring, both the modeled LUMP (\( \beta = -1.33 \)) and LSTM (\( \beta = -1.94 \)) scaling coefficients generally agreed with the observed anti-persistent Brownian noise.

We also investigated the time fractal processes of the flow components that comprise total discharge: quick, intermediate, and slow (Fig. 7 and Table 2). In the process-based South Elkhorn model (SWAT), the quick, intermediate, and slow flow signals (\( \beta = -0.18, -1.19, \) and \(-2.92 \), respectively) became progressively more organized, beginning at near-Gaussian noise, then to anti-persistent Brownian noise, and lastly to persistent Brownian noise. Compared to the SWAT outputs, the process-based Royal Spring model (LUMP) showed greater coherence for all three pathways (\( \beta = -1.08, -2.76, \) and \(-3.75 \), respectively). In a somewhat unexpected result, the corresponding LSTM models had
Intermediate flow. We would expect the opposite relationship to hold. This result could be because of how the recursive digital filter, which separates the flow paths in the LSTM model, is constructed. While the LSTM + RDF estimates of flow magnitude match those of SWAT and LUMP models (Fig. 5), the filter does not itself represent physical underlying processes (Ladson et al., 2013) and thus may not follow scaling relationships. Though we tested only the Lyne-Hollick RDF, there is potential that other filtering methods could resolve this issue (Li et al., 2014).

4. Discussion

We used process-based and data-driven modeling along with correlation and spectral tools to gain insight into the contribution of multiple hydrologic flow paths to discharge generation in surface and subsurface dominated basins. This section discusses our two major findings: (1) we demonstrate the usefulness of data-driven tools, like LSTM models and RDFs, for approximating hydrologic flow pathway contribution and (2) we discuss the particular suitability of LSTM models in settings where subsurface drainage dominates daily-to-monthly water fluxes.


In this study, we assessed the utility of deep learning models to simulate hydrologic flow path contributions in two neighboring systems, a surface-dominated fluvial stream and a subsurface-dominated karst spring, using the same input data. The contrasting hydrologic controls of the two systems allowed for broader inference on flow path behavior. We found that the LSTM models outperformed the process-based SWAT and LUMP models in simulating total flow (Fig. 4). The improved performance was more evident in the groundwater-fed Royal Spring system than in the fluvial South Elkhorn creek. Surprisingly, this improved performance was achieved by using — at most — 25 % of the time series for model training, much less than the typical 70 % split (Kratzert et al., 2018). Our results contrast recent findings that LSTM network models need more training data than hydrologic models to obtain equivalent performance (Bai et al., 2021). In Bai and others (2021), the authors integrate 278 basins from the Model Parameter Estimation Experiment (MOPEX), which are much larger (median: 2358 km², range: 135 to 10,329 km²) than our study basins (~60 km²). Because our basins are smaller, local heterogeneities — if not properly accounted for — may impact process-based model performance more drastically than in larger basins, which smooth heterogeneities over larger spatial scales. Hence, the process-based models in our study might slightly underperform due to unresolved physical processes that the deep learning model does not need to account for explicitly.

Regarding the contribution of hydrologic flow pathways, we found that the LSTM models coupled to a Lyne-Hollick RDF were able to match the general magnitude of quick, intermediate, and slow flow contributions, but inadequately represented the temporal structure of the pathways. Despite the RDF lacking a physical basis (Ladson et al., 2013), its predictions matched closely the process-based SWAT/LUMP predictions ($\beta$ between the model types ranged from 0.58 to 0.71; Fig. 5). The LSTM + RDF predictions struggled most with matching SWAT/LUMP quick flow contributions. This disagreement is likely a function of how the LH filter is constructed whereby increases in overall discharge propagate as increases (of varying magnitude) for each of the three modeled flow paths. In reality, time-varying physical watershed conditions, such as soil moisture, vegetation density, and land cover, could convert the same precipitation input into predominantly quick flow under one set of conditions or primarily intermediate flow under another. Likewise, a flow pathway may become deactivated despite the overall discharge increasing, such as in the case of runoff (quick) ceasing, but subsurface flow (inter) gaining in a stream. With respect to the models, the LSTM simulations consistently over-estimated the auto-correlation of quick flow (Table 1), predicting longer durations of contribution than did the

### Table 2

Power-spectral scaling exponent ($\beta$) for observed and modeled flow and pathway contributions. Values shown are the median with 95% prediction interval in brackets.

<table>
<thead>
<tr>
<th></th>
<th>Power-spectral exponent ($\beta$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Data</td>
</tr>
<tr>
<td>South Elkhorn</td>
<td>Total</td>
</tr>
<tr>
<td></td>
<td>Quick</td>
</tr>
<tr>
<td></td>
<td>Inter</td>
</tr>
<tr>
<td></td>
<td>Slow</td>
</tr>
<tr>
<td>Royal Spring</td>
<td>Total</td>
</tr>
<tr>
<td></td>
<td>Quick</td>
</tr>
<tr>
<td></td>
<td>Inter</td>
</tr>
<tr>
<td></td>
<td>Slow</td>
</tr>
</tbody>
</table>

Fig. 7. Power spectral density plots of data observations and SWAT, LUMP, and LSTM modeling results for total discharge (top row) and hydrologic pathways (bottom two rows). Solid lines represent median results while shaded areas represent the 95% prediction interval.
SWAT/LUMP simulations. However, it is difficult to assess whether LSTM or SWAT/LUMP flow path estimates are “more correct” without additional tracers that can generate pathway contributions that are independent of flow, such as isotopic (e.g., $\delta^2$H$_{2}$O and $\delta^{18}$O$_{2}$O; Klaus and McDonnell 2013) and chemical (e.g., specific conductance; Cartwright and Miller 2021) mass balance mixing models.

In this study, we demonstrated the usefulness of data-driven and statistical tools, like LSTM models and RDPs, for approximating hydrologic flow pathway contribution. Despite the stated limitations, these tools provide a useful first-approximation of hydrologic fluxes from various components within a watershed without the need for extensive model parameterization and calibration, as is the case for traditional process-based numerical models (Moriarti et al., 2012). A few additional points of consideration are how the LSTM and RDP models are themselves constructed. The RDP requires that a dimensionless parameter ($a$) be selected a priori, which can be arbitrary unless other data are available to ascertain the value of $a$ (Li et al., 2014). Further, the number of forward and backward passes to separate flow compartments must be selected, which can also be informed by other modeling estimates (Ladson et al., 2013). In our case, we use our process-based numerical model to inform the likely number of passes. Further, the LSTM network requires the a priori selection of a set of hyperparameter values (Bai et al., 2021). We select a combination of default values from the MATLAB Deep Learning Toolbox and those published in previous work in a variety of hydrologic settings. Notwithstanding these uncertainties, the presented LSTM + RDP model in this study provides useful modeling estimates that can be used in understanding water budgets and the timing, extent, and connectivity of catchment fluxes.

### 4.2. Fluvial versus groundwater control of hydrologic pathways

The degree to which subsurface processes control the overall flow time series was a strong determinant in the ability of LSTM models to achieve satisfactory model performance. In the fluvial South Elkhorn, the LSTM model ($\sigma = 2.30 \text{ m}^2 \text{s}^{-1}$) was not able to match the variability in the data observations ($\sigma = 3.08 \text{ m}^2 \text{s}^{-1}$) or SWAT simulations ($\sigma = 3.05 \text{ m}^2 \text{s}^{-1}$) (Fig. 4). For South Elkhorn, peak discharge is primarily driven by the amount and intensity of rainfall with larger peak rainfall generating larger extreme discharge values (Al Aamery et al., 2018). These extreme flow rates constitute a small fraction of the time series duration and are often several orders of magnitude higher than baseflow. Because extreme events are infrequent, the LSTM learning network is not able to adequately train to such infrequent extremes which results in an underestimation in the range of simulated flow values. This limitation notwithstanding, the LSTM model still outperforms the SWAT model as its improved simulation of discharges within the inner quartile range makes up for its underestimation of peak flows.

On the other hand, in the subsurface-controlled Royal Spring, both LSTM ($\sigma = 0.70 \text{ m}^2 \text{s}^{-1}$) and LUMP ($\sigma = 0.71 \text{ m}^2 \text{s}^{-1}$) model simulations matched the variability in data observations ($\sigma = 0.78 \text{ m}^2 \text{s}^{-1}$). Because Royal Spring is groundwater-fed, there exists an ‘upper-limit’ to the maximum discharge that can be routed through underground conduits (Bonacci, 2001; Husic et al., 2017a). Initial rainfall extremes are attenuated – rather than immediately conveyed to the spring – by the dynamic surficial epikarst and deeper matrix storages, which put a soft upper limit on peak discharge. Once this upper limit to peak discharge is reached, flow in the karst aquifer finds alternate, energetically-favorable routes of transport, including by overtopping sinkholes in surface streams or by overflowing out of secondary springs. The greater attenuation of the rainfall input ($\beta = -0.20$) at the monthly-to-daily hydrologic regime by Royal Spring ($\beta = -1.45$) compared to South Elkhorn ($\beta = -0.92$) is also evident in the time fractal scaling of the basins (Fig. 7). Because the potential values of discharge in Royal Spring are clustered about the mean more closely, the data used for training the LSTM network is sufficient enough for simulating the full range of potential discharges. This limited variability in the range of discharges seems particularly suited to estimation by LSTM modeling frameworks.

The findings of our study suggest that LSTM models are particularly suited to karst systems and potentially to other groundwater-dominated systems. In these basins, precipitation inputs are highly attenuated by internal water stores, an upper limit to peak discharge may exist, and there exists a longer ‘memory’ of previous system states due to gradual export of stored water. These conditions are ideal for training a neural network. While LSTM models have rapidly become adopted by surface water hydrologists for studying runoff generation and streamflow (Bai et al., 2021; Fan et al., 2020; Klotz et al., 2022; Kratzert et al., 2018; Mao et al., 2021), their application to karst systems is less extensive with the authors finding only a few studies in the literature (An et al., 2020; Cheng et al., 2021). Thus, there is a yet unrealized potential to simulate, assess, and understand karst groundwater dynamics using robust LSTM modeling tools.

### 5. Conclusions

In this work, we compared hydrologic pathway contributions from physically-based (SWAT), conceptually-based (LUMP), and deep learning (LSTM) models in a fluvial stream and a karst spring over a twenty-year period. The main findings were:

1. All model types performed satisfactorily, but the LSTM model outperformed both the SWAT and LUMP models in simulating total daily discharge.
2. The LSTM model, coupled with a recursive digital filter, was able to successfully match the magnitude of process-based estimates of quick, intermediate, and slow flow contributions for both basins.
3. The process-based models exhibited more realistic time-fractal scaling of hydrologic flow pathways compared to the LSTM models.

We demonstrate the utility and potential extraction of physical-analogues of LSTM modeling, which will be useful as deep learning approaches to hydrologic modeling become more prominent and modellers look for ways to infer physical information from data-driven predictions.

**CRediT authorship contribution statement**

**Admin Husic:** Conceptualization, Investigation, Methodology, Software, Data curation, Formal analysis, Visualization, Writing – original draft, Writing – review & editing. **Nabil Al-Aamery:** Investigation, Data curation, Writing – original draft, Writing – review & editing. **James F. Fox:** Writing – original draft, Writing – review & editing.

**Declaration of Competing Interest**

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

**Data availability**

The data are stored in an online repository linked in the Acknowledgements.

**Acknowledgments**

We thank an associate editor and two anonymous reviewers for their thorough review of our manuscript. All input data, code, and model results are stored and publicly available via the Open Science Framework (at: http://www.doi.org/10.17605/OSF.IO/CZW32). This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The authors declare no
competing financial or personal interests.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jhydrol.2022.10134.

References


