



In-stream *Escherichia coli* modeling using high-temporal-resolution data with deep learning and process-based models

Ather Abbas¹, Sangsoo Baek¹, Norbert Silvera², Bounsamay Soulileuth³, Yakov Pachepsky⁴, Olivier Ribolzi⁵, Laurie Boithias⁵, and Kyung Hwa Cho¹

¹School of Urban and Environmental Engineering, Ulsan National Institute of Science and Technology, Ulsan 689-798, Republic of Korea

²Institute of Ecology and Environmental Sciences of Paris (iEES-Paris), Sorbonne Université, Univ. Paris Est Creteil, IRD, CNRS, INRA, Paris, France

³IRD, iEES-Paris UMR 242, c/o National Agriculture and Forestry Research Institute, Vientiane, Lao PDR

⁴Environmental Microbial and Food Safety Laboratory, USDA-ARS, Beltsville, MD, USA

⁵Géosciences Environnement Toulouse, Université de Toulouse, CNRS, IRD, UPS, Toulouse, France

Correspondence: Kyung Hwa Cho (khcho@unist.ac.kr) and Laurie Boithias (laurie.boithias@get.omp.eu)

Received: 21 February 2021 – Discussion started: 8 April 2021

Revised: 3 September 2021 – Accepted: 15 October 2021 – Published: 6 December 2021

Abstract. Contamination of surface waters with microbiological pollutants is a major concern to public health. Although long-term and high-frequency *Escherichia coli* (*E. coli*) monitoring can help prevent diseases from fecal pathogenic microorganisms, such monitoring is time-consuming and expensive. Process-driven models are an alternative means for estimating concentrations of fecal pathogens. However, process-based modeling still has limitations in improving the model accuracy because of the complexity of relationships among hydrological and environmental variables. With the rise of data availability and computation power, the use of data-driven models is increasing. In this study, we simulated fate and transport of *E. coli* in a 0.6 km² tropical headwater catchment located in the Lao People's Democratic Republic (Lao PDR) using a deep-learning model and a process-based model. The deep learning model was built using the long short-term memory (LSTM) methodology, whereas the process-based model was constructed using the Hydrological Simulation Program–FORTRAN (HSPF). First, we calibrated both models for surface as well as for subsurface flow. Then, we simulated the *E. coli* transport with 6 min time steps with both the HSPF and LSTM models. The LSTM provided accurate results for surface and subsurface flow with 0.51 and 0.64 of the Nash–Sutcliffe efficiency (NSE) values, respectively. In contrast, the NSE values yielded by the HSPF were -0.7

and 0.59 for surface and subsurface flow. The simulated *E. coli* concentrations from LSTM provided the NSE of 0.35, whereas the HSPF gave an unacceptable performance with an NSE value of -3.01 due to the limitations of HSPF in capturing the dynamics of *E. coli* with land-use change. The simulated *E. coli* concentration showed the rise and drop patterns corresponding to annual changes in land use. This study showcases the application of deep-learning-based models as an efficient alternative to process-based models for *E. coli* fate and transport simulation at the catchment scale.

1 Introduction

Contamination of surface waters through microbiological pollutants is a major public health concern (Bain et al., 2014). Worldwide, pathogens tend to wreak havoc on human health because of the diseases they cause, such as diarrhea, resulting in infant mortality. In particular, developing countries are vulnerable to pathogen-related diseases due to the deficit of sanitation facilities (Boithias et al., 2016). *Escherichia coli* (*E. coli*) has been frequently used as an indicator of fecal bacteria because it is easy to culture (Rochelle-Newall et al., 2015). Higher concentrations of *E. coli* in water tend to be linked to fecal pathogenic microorganisms, harmful to human health. Although long-term and high-frequency *E. coli*

monitoring can help prevent waterborne diseases from fecal pathogenic microorganisms, monitoring *E. coli* concentrations is time-consuming and expensive (Cho et al., 2016; Frolich et al., 2017; Kim et al., 2017). Furthermore, high-frequency datasets of *E. coli* concentration are scarce, and available long-term datasets are often inadequate to yield a continuous time series of fecal pathogenic microorganisms (van der Leeuw, 2004). Modeling approaches can overcome this drawback in monitoring. Thus, they can be a means to determine the fate and transport of fecal pathogenic microorganisms at the catchment scale by simulating *E. coli* in environmental compartments, such as the soil surface and streams (Ligaray et al., 2016; Pachepsky and Shelton, 2011).

Several process-based models have been developed to simulate stream water contamination by *E. coli*. Popular models to simulate *E. coli* are the Soil and Water Assessment Tool (SWAT) (Neitsch et al., 2011), Hydrological Simulation Program–FORTRAN (HSPF) (Bicknell et al., 1997), INCA-Pathogens (Whitehead et al., 2016), and pathogen catchment budget (PCB) (Ferguson et al., 2007). The fate and transport of *E. coli* are a complex phenomenon with several drivers (Pachepsky et al., 2018), such as the hydrological regime (Boithias et al., 2016; Pachepsky et al., 2017), contributions of both surface runoff and subsurface flow to the overall in-stream discharge (Boithias et al., 2021b), concentration and sources of suspended sediment (Ribolzi et al., 2016; Nguyen et al., 2016), land use (Causse et al., 2015; Nakhle et al., 2021), intrinsic properties of the bacterium (Pachepsky et al., 2014), and economic conditions (Iqbal et al., 2019). Recently, Sowah et al. (2020) applied the SWAT model to research the sources and drivers of *E. coli* in the Clouds Creek watershed, USA. However, the process-based models still have limitations to accuracy due to the complexity of relationships among hydrological and environmental variables (Abimbola et al., 2020). In addition, the simplified equations of these models can increase the inherent uncertainties, resulting in simulation errors. To overcome these limitations, several modifications of the *E. coli* module of the SWAT model have been proposed to incorporate the impacts of the multiple drivers of *E. coli* fate and transport (Kim et al., 2018; Meshesha et al., 2020). The *E. coli* concentration in surface water varies significantly within a very short time period (Chen et al., 2014; Boithias et al., 2021b). Daily simulations cannot capture the dynamics of *E. coli* in a short duration. In particular, the simulation with high temporal resolution is important in small headwater catchments because the duration of flood events might be less than 1 d (Gassman et al., 2007). Therefore, an *E. coli* concentration simulation with high temporal resolution should be conducted to determine the temporal distribution of *E. coli*.

Recently, deep learning (DL) has become a promising alternative approach for estimating water quality by using features of water constituent dynamics (Pyo et al., 2021). Deep-learning-based models are superior to their process-based counterparts due to their high accuracy, faster pre-

diction time, and ability to model complex physical phenomena (Sze et al., 2017). Deep-learning models can exploit a particular compositionality in the input features by finding more abstract features in them (Bengio et al., 2021). Long short-term memory (LSTM) networks have an advantage over other deep-learning-based models in that they can extract complex patterns from sequence data (Schmidhuber and Hochreiter, 1997). Several studies have applied deep learning to water quality modeling and prediction (Peterson et al., 2020; Isikdogan et al., 2017; Solanki et al., 2015). Dong et al. (2019) used LSTM to predict dissolved oxygen concentrations and showed that LSTM performs better than other machine-learning methods, such as autoregressive integrated moving average or artificial neural networks. Although LSTM has been used extensively for building hydrological models (Abbas et al., 2020), its potential has not yet been explored to estimate *E. coli* concentration in stream waters. Deep-learning-based models have also not been developed for the simulation of water quality with high temporal resolution.

This study aims to evaluate the applicability of LSTM to simulate in-stream *E. coli* concentrations with high temporal resolution. In addition, the process-based model HSPF was used as a benchmark to compare and assess the performance of LSTM. Both models were applied to a 0.6 km² tropical headwater catchment from the northern Lao People's Democratic Republic (PDR). The temporal resolution of the simulations was 6 min in both models. Thus, the specific objectives of this study were to compare the performance of a process-based model and a deep-learning model (1) to simulate both surface and subsurface flow, (2) to simulate *E. coli* concentration, and (3) to analyze the response of *E. coli* to changing land use.

2 Materials and methods

2.1 Study site and data acquisition

The study area is the 0.6 km² Houay Pano headwater catchment, located 10 km south of the city of Luang Prabang, Lao PDR (Boithias et al., 2021a) (Fig. 1). This catchment is representative of montane agroecosystems in Southeast Asia and is part of the long-term critical zone observatory network called multiscale TROPical CatchmentS (M-TROPICS), which is affiliated with the French research infrastructure OZCAR (Gaillardet et al., 2018). This site had undergone rapid land-use changes from 2011 to 2018 (Fig. S1a). The characteristics of this area, including land-use information, are provided in the Supplement (Sect. S1). We collected weather, hydrological, *E. coli* concentration, and electrical conductivity data at 6 min time steps from 2011 to 2018. Rainfall, relative humidity, solar radiation, wind speed, and air temperature were measured with an automatic weather station Campbell Scientific BWS200, which was

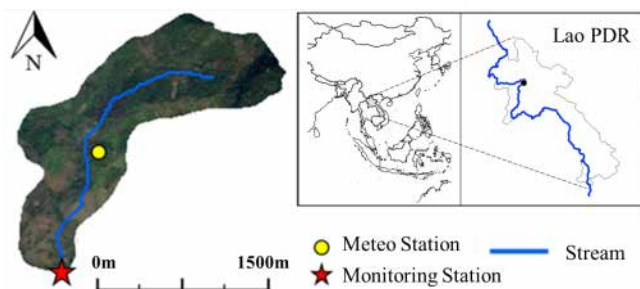


Figure 1. Location of the study area. The study area is located near Luang Prabang in the northern Lao PDR. The gauging and monitoring station is located at the outlet of the catchment, where water level is recorded and where water samples are collected for *E. coli* concentration measurement. Climate data were measured at the meteorological station.

equipped with ARG100 (a 0.2 mm capacity tipping bucket). The potential evapotranspiration was calculated using the Penman–Monteith method. We measured the stream water level at the monitoring station using a V-notch and water-level recorder (OTT Thalimedes). The discharge was estimated based on the rating curve relating discharge to water levels. The surface flow and subsurface flow were calculated using the electrical conductivity method (Ribalzi et al., 2018). A detailed description of this method is provided in the Supplement (Sect. S2). *E. coli* concentration was measured based on the standardized microplate method (ISO 9308–3). A detailed explanation of the *E. coli* experiment can be found in the Supplement (Sect. S3). In this study, we carried out biweekly grab sampling of *E. coli* from 2011 to 2018. Over the same period, we also monitored 11 flood events to assess *E. coli* dynamics during flood events using an automated sampler (ICRISAT) triggered by the water-level recorder to collect water after every 2 cm water-level change during flood rising and every 5 cm water-level change during flood recession. The total number of *E. coli* samples collected over the 2011–2018 period was 255. In addition, we collected the monthly numbers of poultry, swine, goats, and the number of people who visited the study area. These data were used to quantify the source of *E. coli* in this catchment (Rochelle-Newall et al., 2016) (Fig. S1b).

2.2 Flow and *E. coli* concentration simulation

HSPF and LSTM models were used to simulate in-stream surface flow, subsurface flow, and *E. coli* concentration. HSPF and LSTM are popular models (Bicknell et al., 1997; Ahmadisharaf and Benham, 2020; Kratzert et al., 2019). Both models have been used for hydrological and water quality simulations (Peterson et al., 2020; Isikdogan et al., 2017; Ahmed et al., 2014). In the HSPF, the simulation of surface and subsurface flow and of *E. coli* concentration was carried out in three steps: (1) building the model, (2) conducting sensitivity analysis based on Latin hypercube–one factor

at a time (LH-OAT), and (3) calibrating the model using the Newton algorithm (Nash, 1984). The schematic of the LSTM simulation is shown in Fig. 2. The first step in building this model was data preparation (Fig. 2a). LSTM then simulated surface and subsurface flow with weather data (Fig. 2b). Finally, we estimated the *E. coli* concentration at 6 min intervals using rainfall, bacteria source, land-use change, and surface and subsurface flow (Fig. 2c). The fecal matter from the *E. coli* sources was assumed to be evenly distributed in the catchment. The monthly *E. coli* source data are presented in Fig. S1b. The time-series data of the *E. coli* source were used as input for the *E. coli* simulation.

2.2.1 HSPF

The HSPF model is a process-driven model that simulates processes at the catchment scale (Bicknell et al., 1997). It has been extensively used to model the fate and transport of *E. coli* in catchments (Ahmadisharaf and Benham, 2020; Chin et al., 2009) and to develop total maximum daily loads of *E. coli* at various locations (Mishra et al., 2018; Yagow et al., 1998). The original software was written in the FORTRAN programming language. The Hydrological Simulation Program Python (HSP2) was recently developed based on the Python programming language (van Rossum, 2007). HSP2 is a platform-independent software that extends the functionality of HSPF by allowing the use of dynamic variables and easier management of input and output files (Heaphy et al., 2015). The HSPF simulates the hydrological regimes by discretizing the catchment into pervious and impervious hydrological response units (HRUs). Evapotranspiration, surface retention, surface infiltration, interflow, baseflow, and deep percolation are simulated at pervious HRUs, whereas surface retention and surface flow are simulated at impervious HRUs (Bicknell et al., 1997). The simulation of in-stream *E. coli* concentration in HSPF is based on a first-order kinetics approach, considering the decay rate (Fonseca et al., 2014). Detailed descriptions of hydrological and *E. coli* simulations can be found in Bicknell et al. (1997). For this study, we converted the original FORTRAN code of the *E. coli* module of HSPF into the Python programming language. This allowed us to incorporate more efficient use of input data, such as the annual change in land use and the monthly bacterial source.

In our study, HRUs were divided into four groups based on land use: “Forest”, “Fallow”, “Teak”, and “Annual” crops. Among land uses, we did not consider any imperviousness in Forest and Fallow. We considered 2 % and 1 % imperviousness for the Teak and Annual crop land uses (Patin et al., 2018). We selected 13 and 4 parameters for each land use for the sensitivity analysis of hydrological and *E. coli* simulations, respectively (Tables 1 and S1). The total number of parameters for hydrological and *E. coli* simulation were 52 and 18, respectively. In model calibration, we selected the 25 most sensitive parameters of the hydrological simulation and all parameters of the *E. coli* simulation. Sensitivity anal-

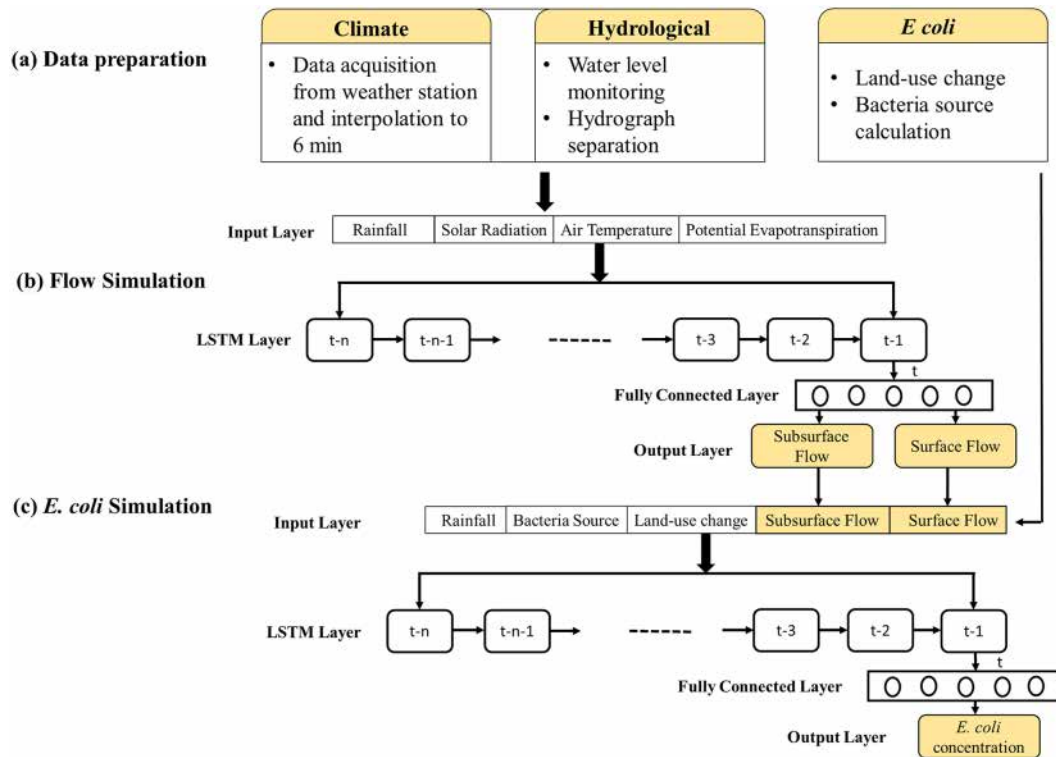


Figure 2. Structure of the LSTM model. Environmental data are used to predict surface flow and subsurface flow. Simulated flows along with bacteria source, land-use information, and rainfall are used to simulate *E. coli* concentration. The “ n ” represents the length of input data used by LSTM.

ysis and model calibration were conducted based on the LH-OAT and the Newton algorithm, respectively. A detailed explanation of the LH-OAT and the Newton algorithm can be found in the Supplement (Sect. S4).

2.2.2 LSTM

In the data preparation step (Fig. 2a), our data were converted to the 6 min frequency. This was carried out by interpolating the hourly weather data. Rainfall data were already available at 6 min for 2011 and 2012, while for 2013 to 2018 they were available at 1 min frequency and were aggregated into a 6 min time series. For *E. coli* concentration, the values nearest to a 6 min step were used as representative of that time step. We then built the LSTM model to simulate surface and subsurface flow using the validated model structure (Abbas et al., 2020) (Fig. 2b). It uses historical data of rainfall, solar radiation, air temperature, and potential evapotranspiration to simulate surface and subsurface flow. To simulate the output at a time step “ t ”, LSTM uses the data of previous “ n ” time steps as inputs (Chollet, 2017). The inputs from previous “ n ” time steps are used by LSTM to predict the output at the next time step “ $(t + 1)$ ”. These time steps are called “lookback” steps (Chollet, 2017). The simulated surface and subsurface flows from the LSTM were applied to simulate the *E. coli* concentration (Fig. 2c). We adopted a bacterial source and

land-use information as an input for the LSTM. The preprocessing of the data before feeding the neural network can have a significant impact on the performance (Banhatti and Deka, 2016). Therefore, we compared the performance of the model by transforming the *E. coli* concentration using the min–max transformation and the logarithmic transformation. The min–max transformation results in data between 0 and 1, while logarithmic transformation transforms the data on a logarithmic scale. To investigate the impact of land-use change on in-stream *E. coli* concentrations, we conducted *E. coli* simulations in two scenarios. In scenario 1, we used the land-use change and *E. coli* source information separately. In scenario 2, the number of input features was reduced by multiplying the *E. coli* source by land-use change. In this way, we calculated the *E. coli* source per area for each land use and used this as input instead of using land use and *E. coli* information as separate input features.

LSTM is a special recurrent neural network designed to extract temporal features from sequence data (Hochreiter and Schmidhuber, 1997). An LSTM cell is the basic building block of the LSTM (Fig. S2). It consists of three “gates” and two “states”. The gates are “forget”, “update”, and “output”, which decide what information to forget, allow in, and allow out from the LSTM “memory”, respectively. The states act as

Table 1. Optimal values and range of HSPF parameters for surface and subsurface flows and *E. coli* concentration. Bold parameters were optimized during the flow calibration process. All parameters related to *E. coli* were optimized during model calibration.

Parameters		Land use				Lower Limit	Upper Limit
		Forest	Teak	Fallow	Annual Crop		
Surface and subsurface flow	INFILT	0.31	0.39	0.39	0.36	0.001	0.5
	INFILD	2.0	1.94	1.95	1.55	1.0	3.0
	INTFW	2.60	7.01	7.01	5.64	1.0	10.0
	UZSN	1.36	1.47	0.84	1.24	0.05	2.0
	LZSN	8.88	9.43	4.18	8.66	2.0	10.0
	AGWETP	0.02	0.007	0.02	0.06	0.0	0.2
	NSUR	0.18	0.39	0.15	0.43	0.05	0.5
	BASETP	0.05	0.09	0.095	0.003	0.0	0.2
DEEPR	0.28	0.16	0.21	0.20	0.0	0.5	
<i>E. coli</i> concentration	SQOLIM MF	4.99	1.35	2.04	0.53	0.5	10
	WSQOP	9.12	9.31	8.87	9.38	0.1	10.0
	IOQC	5367	8337	8380	8756	1000	10 000
	AOQC	8672	7474	5465	8776	1000	10 000
	FSTDEC			3.04		0.1	10.0
	THFST			1.92		1.01	2.0

a memory or information carrier across time. The equations describing the functions of gates and states are as follows:

$$C_c^{<t>} = \tanh(W_c[h^{<t-1>}, x^{<t>}] + b_c), \tag{1}$$

$$\Gamma_f = \sigma(W_f[c^{<t-1>}, x^{<t>}] + b_f), \tag{2}$$

$$\Gamma_o = \sigma(W_o[c^{<t-1>}, x^{<t>}] + b_o), \tag{3}$$

$$\Gamma_u = \sigma(W_u[c^{<t-1>}, x^{<t>}] + b_u), \tag{4}$$

$$C^{<t>} = \Gamma_u * C_c^{<t>} + \Gamma_f * C^{<t-1>}, \tag{5}$$

$$h^{<t>} = \Gamma_o * \tanh C^{<t>}. \tag{6}$$

The symbol * in the above equations represents element-wise multiplication. The behavior of each gate is controlled by the weights (*W*) and biases (*b*) associated with them. Gate output is further modified by a nonlinear function (σ). At each time step (*t*), the prospective cell state ($C_c^{<t>}$) is calculated based on the output from the previous time step ($h^{<t-1>}$) and the input from the current time step ($x^{<t>}$) (Eq. 1). The notation $W_c[h^{<t-1>}, x^{<t>}]$ represents point-wise multiplication of new inputs and previous hidden state with the weight matrix W_c and then adds their output. This prospective cell state ($C_c^{<t>}$), along with the output from the “forget” and “update” gates, decides the current cell state ($c^{<t>}$) (Eq. 5). The current cell state and output gate control the output values from LSTM ($h^{<t>}$) forming the so-called hidden state (Eq. 6). The hyperbolic tangent (\tanh) is another nonlinearity used in LSTM for the calculation of the cell state (Eq. 1) and the output state (Eq. 6). Equations (1)–(6) are used to calculate the LSTM output, which is then compared

with observed values to calculate the error. This study used the mean square error (MSE) as the error metric.

We used the TensorFlow software v1.15 for building the LSTM model (Abadi et al., 2016). We used an Intel® Core™ i7-9700 processor with a graphics card of NVIDIA GeForce RTX 2080 with 12 GB of dedicated GPU memory, along with 64 GB of random-access memory for simulating surface, subsurface, and *E. coli*.

2.2.3 Hyperparameters of LSTM

The structure and performance of the LSTM were controlled by hyperparameters, including the dropout rate, the number of LSTM units, learning rate, lookback steps, and activation functions for both LSTM and the fully connected layer (Table 2). Dropout is a regularization technique that switches off a certain number of nodes in the LSTM (Goodfellow et al., 2016). This simple technique helps break the brittle coadaptation of weights, which hinders generalization to unseen data. This way, dropout prevents overfitting (Srivastava et al., 2014). In the case of overfitting, the model performs better on calibration data, but its performance deteriorates on new unseen data. The number of LSTM units directly corresponds to the learning capacity of LSTM, but it also accounts for more memory and computation. This number determines the size of the weight matrix of an LSTM. The learning rate defines the change in the weights of the neural network during calibration (Goodfellow et al., 2016). A higher number of lookback steps allows LSTM to capture long-term patterns at the cost of an increase in memory consumption and computation. The activation function determines the nonlinearity in the model.

Table 2. Hyperparameters of LSTM for surface flow, subsurface flow, and *E. coli* concentration simulation.

Parameter	Surface and subsurface flow	<i>E. coli</i>
Activation function (LSTM layer)	Rectified linear unit (ReLU)	Rectified linear unit (ReLU)
Activation function (dense layer)	Rectified linear unit (ReLU)	Rectified linear unit (ReLU)
Batch size	128	16
Learning rate	1×10^{-5}	1×10^{-6}
lookback steps	5 h	5 h
Dropout	0.3	0.3
Hidden units	64	100
Input data	Rainfall, solar radiation, air temperature, potential evapotranspiration	Rainfall, surface flow, subsurface flow, land use, bacteria source
Calibration epochs	500	7000
Training samples	490 000	182
Test samples	210 000	73

2.3 Performance statistics

Evaluations to assess the performance of the HSPF and LSTM were conducted using MSE, Nash–Sutcliffe efficiency (NSE), and percent bias (PBIAS) (Nash and Sutcliffe, 1970; Gupta et al., 1999). NSE is useful for interpreting the model performance by generating a dimensionless value as the performance index (Lin et al., 2017). The PBIAS measures the average tendency of the simulated data to be overestimated or underestimated than observed values (Moriassi et al., 2015). The MSE, NSE, and PBIAS were calculated using the following equations:

$$\text{MSE} = \frac{[\sum_{i=1}^m (o_i - p_i)^2]}{m}, \quad (7)$$

$$\text{NSE} = 1 - \frac{\sum (o_i - p_i)^2}{\sum (o_i - \bar{o})^2}, \quad (8)$$

$$\text{PBIAS} = 1 - \frac{\sum_{i=1}^m (o_i - p_i)}{\sum_{i=1}^m o_i}, \quad (9)$$

where p_i is the simulated data, o_i is the observed data, and m is the number of points in the data.

3 Results and discussion

3.1 Land-use change and *E. coli* source

The land-use change from 2011 to 2018 is shown in Fig. S1a. The area of Fallow land use increased from 2011 to 2016, whereas Annual crop area decreased. Teak tree plantations were expanded until 2013 and were retained. Forest land use accounted for about 10 % of the study area from 2011

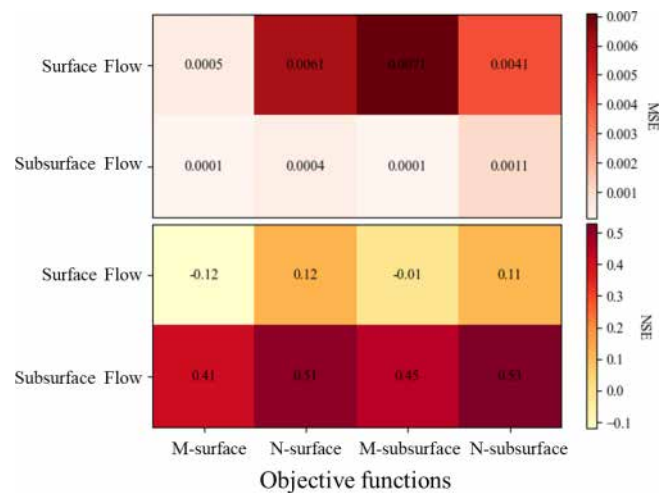


Figure 3. Performance of the HSPF model with different objective functions (e.g., M-surface, N-surface, M-subsurface, and N-subsurface). The color indicates the value of MSE and NSE. M-surface is the objective function based on MSE and surface flow, N-surface is the objective function based on NSE and surface flow, M-subsurface is the objective function based on MSE and subsurface flow, and N-subsurface is the objective function based on NSE and subsurface flow.

to 2018. In general, the land-use change was dynamic from 2011 to 2013, whereas its variation diminished from 2016 to 2018. Previous studies have demonstrated that the expansion of Teak trees might increase the surface flow (Ribolzi et al., 2017; Song et al., 2020). Higher runoff at the soil surface may cause a higher inflow of *E. coli* with surface flow. The monthly *E. coli* source in the catchment decreased from

2×10^{15} in 2011 to 3×10^{14} in 2018 (Fig. S1b). This decrease in the *E. coli* source is caused by the decrease in manpower needed in Teak tree plantations and in Fallow plots compared to the Annual crop (Fig. S1a) (Boithias et al., 2021b).

3.2 Sensitivity analysis and optimization result

The sensitivity results for the flow simulation are shown in Fig. S3, and the most sensitive parameters are listed in Table S2. The interflow and infiltration-related parameters were the most sensitive parameters for surface and subsurface flows. Manning's "n" value (NSUR) for Teak and Fallow land uses was among the 10 most sensitive parameters. Kim et al. (2017) suggested that Manning's coefficient value is the most sensitive parameter in the hydrological simulation of tropical headwater catchments, such as the Houay Pano catchment in the northern Lao PDR. The groundwater recession rate (AGWRC) and soil infiltration capacity (INFILT) were sensitive to subsurface flow. In Annual crop land use, infiltration capacity (INFILT) and upper zone storage (UZSN) were the most sensitive parameters. Abbas et al. (2020) demonstrated that INFILT is the most sensitive parameter for subsurface flow in tropical subcatchments.

The sensitivity analysis results for *E. coli* are shown in Fig. S4 and Table S3. The parameters related to the transport of *E. coli* on the land surface (e.g., WSQOP, SQOLIM_MF) were more sensitive than other parameters. IOQC and AOQC were the least sensitive parameters. These parameters are related to *E. coli* transport in interflow and baseflow (Bicknell et al., 1997). This implies that the in-stream *E. coli* concentration at the study site is mainly driven by surface flow (Boithias et al., 2021b). A previous study also demonstrated that 89% of in-stream *E. coli* concentrations were driven by surface flow (Boithias et al., 2021b). Figure 3 shows the model performance dependent on different objective functions. We found that the model performance was better when the NSE was selected as the objective function. The NSE of the surface and subsurface flow was positive by optimizing with NSE. However, the NSE value for surface flow was negative when the objective function was MSE during the optimization. Negative NSE indicated an "unsatisfactory" performance range (Moriassi et al., 2015).

3.3 Flow simulation

The simulated surface and subsurface flows using the HSPF are plotted in Fig. 4. We found that the simulated subsurface flow was underestimated compared to the observations. Although surface flow from the HSPF followed the trend and peaks of observations, this model yielded a negative NSE value, indicating that the model simulation was unacceptable (Moriassi et al., 2015) (Table 3). The NSE values for subsurface flow from HSPF were 0.49 and 0.59 for calibration and validation, respectively. Hence, the HSPF model was better at simulating subsurface flow than surface flow. In particular,

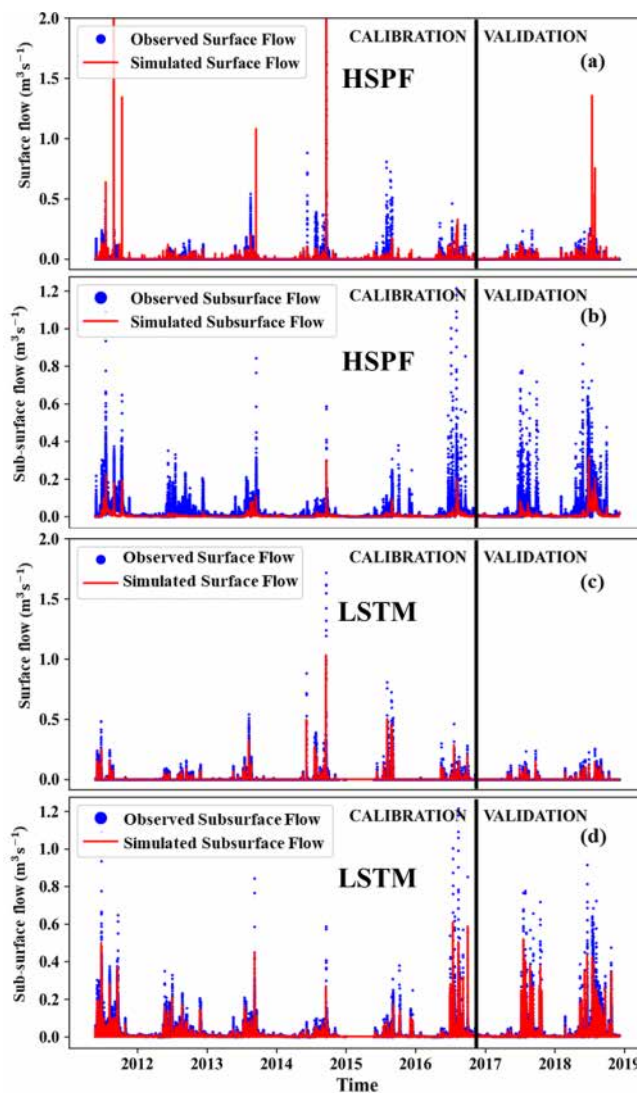


Figure 4. Hydrological simulation from HSPF and LSTM: (a) simulated and observed surface flow from HSPF, (b) simulated and observed subsurface flow from HSPF, (c) simulated and observed surface flow from LSTM, and (d) simulated and observed subsurface flow from LSTM.

the simulated surface flow was underestimated compared to the observations. The average values of INFILT and UZSN were 0.36 and 1.22, respectively, which were larger than those reported in previous studies (Lee et al., 2020). INFILT controls the overall division of available moisture into the surface and subsurface (Bicknell et al., 1997). The parameter UZSN influences the evapotranspiration process (Bicknell et al., 1997). This underestimation of surface flow using HSPF is consistent with a previous study (Kim et al., 2017). We also investigated the impact of underestimation and overestimation of the flow by plotting flow duration curves (Fig. S5). Although both flows can capture the peak flow, the simulated

Table 3. Performance metrics of HSPF and the LSTM model for surface and subsurface flow.

Model	Flow type	Scenario	MSE ($\text{m}^3 \text{s}^{-1}$)	NSE	PBIAS
HSPF	Surface flow	Calibration	6.4×10^{-4}	-0.02	-59
		Validation	4.7×10^{-5}	-0.7	-28
	Subsurface flow	Calibration	2.7×10^{-4}	0.49	-51
		Validation	5×10^{-4}	0.59	-22
LSTM	Surface flow	Calibration	1.4×10^{-4}	0.56	-48
		Validation	1.9×10^{-4}	0.51	-63
	Subsurface flow	Calibration	5.4×10^{-3}	0.69	-42
		Validation	5.9×10^{-3}	0.64	-46

subsurface flow was still underestimated compared to the observed subsurface flow.

The simulated surface and subsurface flows using the LSTM model are plotted in Fig. 4. The NSE values for the calibration period were 0.56 and 0.69 for surface and subsurface flow, respectively. The corresponding validation NSE values of the surface and subsurface flows were 0.51 and 0.64, respectively. These results indicate that the LSTM had a satisfactory performance for both the calibration and validation periods according to the criteria of Moriasi et al. (2015). LSTM overcame the problem of the HSPF model underestimating subsurface flow. In addition, the peak surface flows from the LSTM were similar to observations. The observed and simulated flows in storm events are presented in Figs. S6–S11. The simulated surface flow by HSPF followed the rainfall events more closely as compared to that of LSTM. The peaks in surface flow in Fig. S8 are completely missed by LSTM but captured by the HSPF model. We also noted that LSTM can follow the observed trends in surface and subsurface flow more closely than the HSPF (Figs. S6, S9, S10). The falling limb from the predicted subsurface flow of LSTM is gentle and follows the observed pattern (Figs. S9–S11). This leads to increased NSE values for both surface flow as well as for subsurface flow. The hyperparameters of the LSTM are described in Table 2. The rectified linear unit (ReLU) was chosen as the activation function for the LSTM output. Because the simulated *E. coli* should be positive, we chose ReLU, which cannot produce negative values from the model (Nair and Hinton, 2010). The optimal batch size and LSTM units were 128 and 64, respectively. The optimal value of the lookback steps was 50, which is equal to 5 h of input data.

We analyzed the model performance for surface and subsurface flows during storm events (Fig. 5). The events were selected in which the peak flow exceeded $0.2 \text{ m}^3 \text{ s}^{-1}$. The performance of LSTM is considerably better than that of HSPF for most storm events. In surface flow, the average MSE of LSTM and HSPF was 1.1×10^{-4} and 6.1×10^{-4} ($\text{m}^3 \text{ s}^{-1}$), respectively. The NSE values from LSTM varied from 0.2 to 0.6, whereas those of HSPF ranged

Table 4. Performance metrics of HSPF and LSTM for *E. coli* concentration simulation.

Model	Scenario	MSE (MPN 100 mL^{-1})	NSE	PBIAS
HSPF	Calibration	1.4×10^8	-0.29	-58
	Validation	1.9×10^8	-3.01	73.01
LSTM	Calibration	7.1×10^6	0.39	-1.49
	Validation	3.0×10^7	0.35	62.72

from -1.0 to 0.4. We found that the NSE values from the HSPF vary considerably depending on storm events. On 11 June 2015, the NSE value of HSPF was as high as 0.4, whereas for some other dates it was below 0. Although the subsurface flow of the HSPF provided better model performance than surface flow simulation, this model still presented an unacceptable result, with a negative NSE value.

3.4 *E. coli* simulation

Figure 6 shows the temporal distribution of *E. coli* concentration using HSPF and LSTM. The *E. coli* concentration from HSPF was overestimated. The performance matrices of the HSPF were also worse than those of the LSTM (Table 4). In particular, the HSPF simulation presented a PBIAS value of 73, indicating an overestimation of *E. coli* concentration (Moriasi et al., 2015). Ackerman and Weisman (2014) reported that the *E. coli* from HPSF was overestimated compared to observation. The overestimation of simulated *E. coli* at tropical sites has also been observed by Kim et al. (2017). *E. coli* simulation from LSTM is satisfactory in both calibration and validation periods according to the criteria set by Moriasi et al. (2015). In contrast, the HSPF result can be regarded as “unsatisfactory” in both the calibration and validation periods. These results implied that LSTM could generate acceptable performances and had good agreement between the observed and simulated *E. coli*.

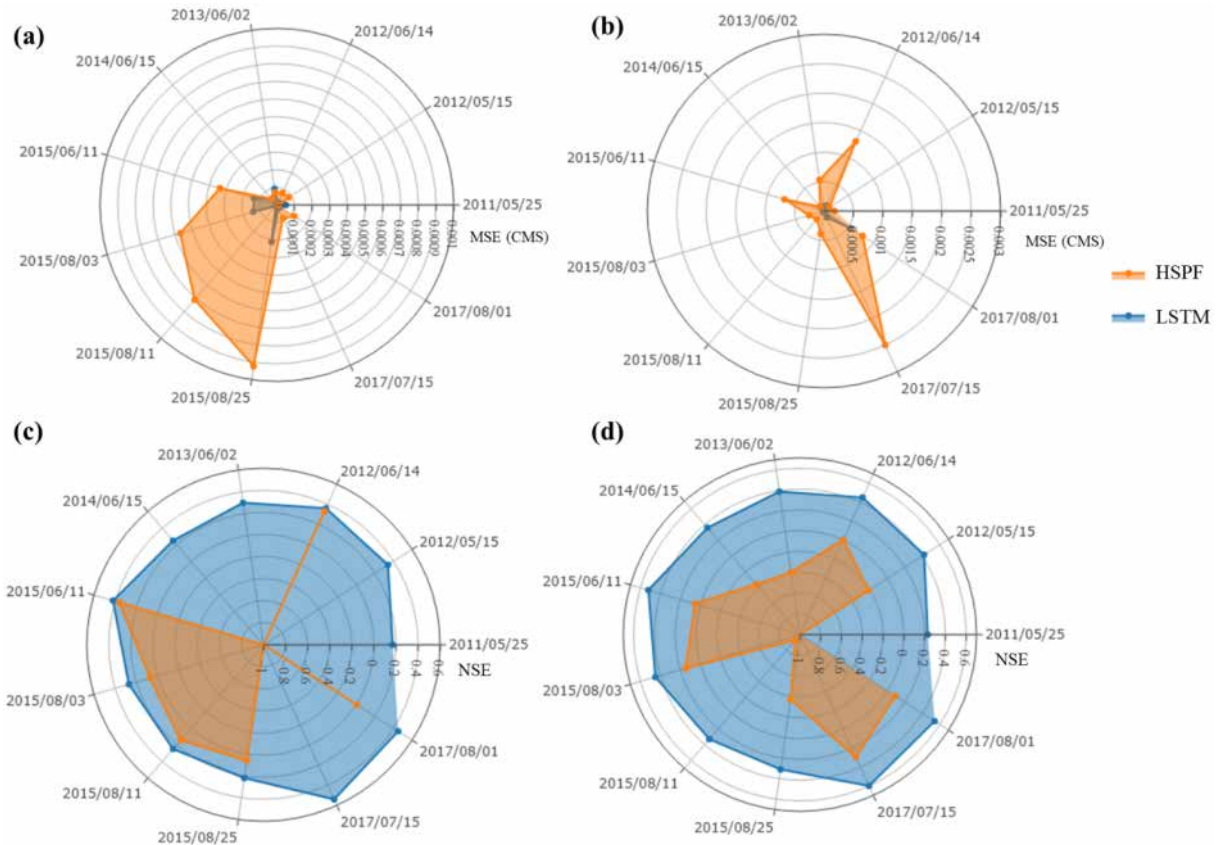


Figure 5. Comparison of the hydrological simulation during storm events: (a) MSE value of the surface flow, (b) MSE value of the subsurface flow, (c) NSE value of the surface flow, and (d) NSE value of the subsurface flow.

The simulation results during the storm events using both the HSPF and LSTM models are shown in Figs. 7 and S6–S11. Figure 8 shows two storm events from the validation data during July and August 2017, whereas the other figures show the storm events from the calibration data. In general, the simulated *E. coli* by HSPF and LSTM were overestimated and underestimated, respectively. This difference might be caused by the fact that *E. coli* from HSPF is more responsive to surface flow, whereas *E. coli* from LSTM is more influenced by subsurface flow (Ackerman and Weisman, 2014). The sensitivity analysis of HSPF also demonstrated that the influence of interflow and baseflow on *E. coli* is weaker than that of surface flow because the parameters IOQC and AOQC are the least sensitive parameters for *E. coli* simulation. Both parameters affect the *E. coli* concentration in interflow and baseflow (Bicknell et al., 1997). The simulated *E. coli* of LSTM rose sharply and dropped slowly, similarly to the observations, whereas that of the HSPF decreased steeply (Figs. S6–S11). Although both models simulated the peak time of the *E. coli* correctly, the HSPF was limited in its ability to simulate the slope of the falling limb. This performance difference between models was caused by the extent of the influence from hydrological variables (e.g.,

rainfall, surface flow, and subsurface flow) on model output. LSTM was effective in reflecting the response of the output to hydrologic variables (Kratzert et al., 2019).

We observed that both HSPF and LSTM simulated peaks even when the observed data did show corresponding peaks (Figs. S8 and S11). The peaks predicted in Fig. S8 are solely from HSPF, while the peak event in Fig. S11 is predicted by both HSPF and LSTM models. This shows the efficacy of both calibrated models. We could conclude from Fig. S11 that the lack of an observed peak is more likely because of missing observation. However, a similar conclusion cannot be drawn for all the predicted *E. coli* peaks in Fig. S8 because of contradicting results of LSTM and HSPF.

The performance metrics for the LSTM and HSPF models during storm events are shown in Fig. 8. In general, we observed better LSTM performance than HSPF in terms of NSE and MSE values. The HSPF model performed better than the LSTM for only two storm events: on 15 June 2014 and 11 June 2015. For the remaining storm events, the NSE values from LSTM are higher than those of the HSPF – an NSE range from 0.20 to 0.65. Similarly, for MSE values, the LSTM was superior to the HSPF for all storm events except for the storm events on 15 June 2014 and 11 June 2015.

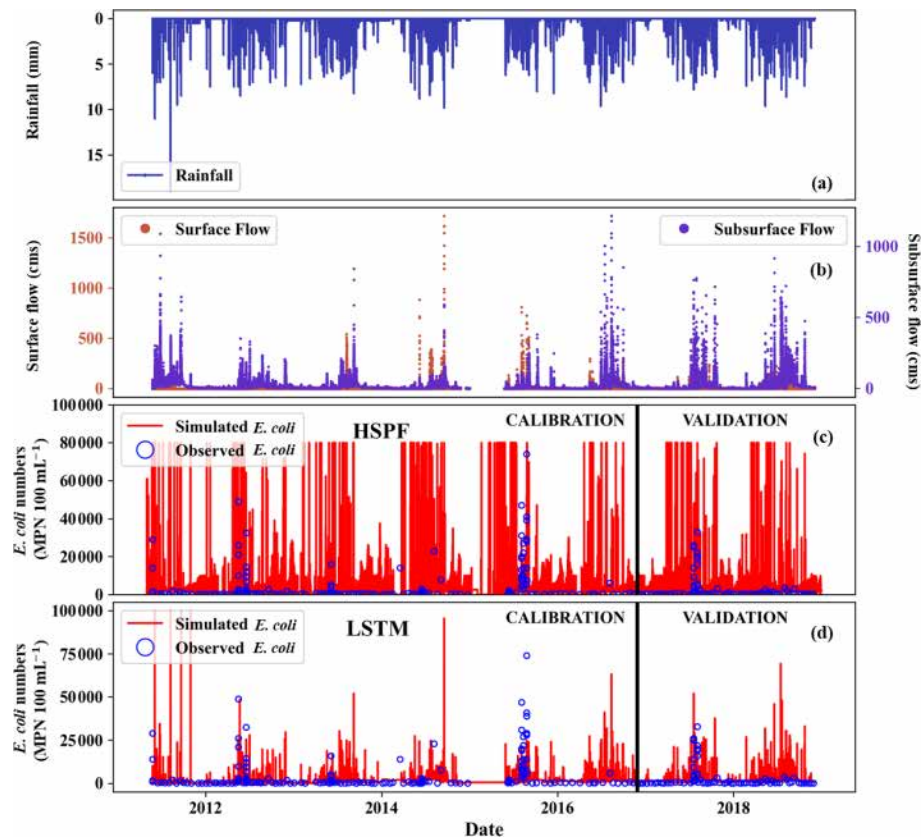


Figure 6. *E. coli* simulation from LSTM and HSPF: (a) measured rainfall, (b) observed surface and subsurface flow, (c) simulated and observed *E. coli* concentration using HSPF, and (d) simulated and observed *E. coli* concentration using LSTM.

We observed the impact of logarithmic and min–max transformations on the model performance (Fig. 9). The results of the logarithmic transformation were closer to the observations than the min–max transformation by showing an NSE of 0.57. A negative PBIAS value was obtained in logarithmic transformation. This indicated that the simulated *E. coli* from logarithmic transformation was underestimated, whereas the result of the min–max transformation was overestimated. This behavior can be attributed to the ability of min–max scaling to be more sensitive to outliers (Chuang et al., 2010). As a result, if a better accuracy during storm events is required, the target variable can be transformed on a logarithmic scale prior to calibration. This is because log transformation can reduce the effect of outliers (Singh and Kingsbury, 2017). It has been reported that log transformation can improve the performance of data-driven models when the data contain outliers (Zheng and Casari, 2018).

3.5 *E. coli* response to land-use change

We investigated the impact of land-use change and bacterial sources on the in-stream *E. coli* concentration simulation (Fig. 10). In scenario 1, we used land-use change time-series information (Fig. S1a) and bacterial source informa-

tion (Fig. S1b). In scenario 2, we divided the bacterial source by the fraction of each land use (Fig. S2c). In scenario 1, we observed a larger variation in *E. coli* concentration from 2014 to 2018 (Fig. 10a), whereas in scenario 2, the variation in *E. coli* was smaller than that in scenario 1 (Fig. 10b). This variation in *E. coli* was due to land-use change in scenario 1. In particular, *E. coli* in 2016 was less than in other years because the Annual crop land use decreased. On the other hand, the variation in *E. coli* was not observed in scenario 2 from 2015 to 2017. Neither scenario showed a significant response from 2011 to 2014. During these years, the rise in Fallow land use was complemented by a decrease in Annual crop land use.

3.6 Limitations and future research

Transport of soil particles by surface flow and suspended sediments within the stream plays a crucial role in the fate and transport of *E. coli* (Thupaki et al., 2013). Several studies have emphasized the importance of particle size (Cho et al., 2010), adsorption to soil and sediment particles (Palmatteer et al., 1993), and resuspension of *E. coli* (Kim et al., 2017) with streambed sediments for modeling the fate and transport of *E. coli* at the catchment scale. In this study, we considered nei-

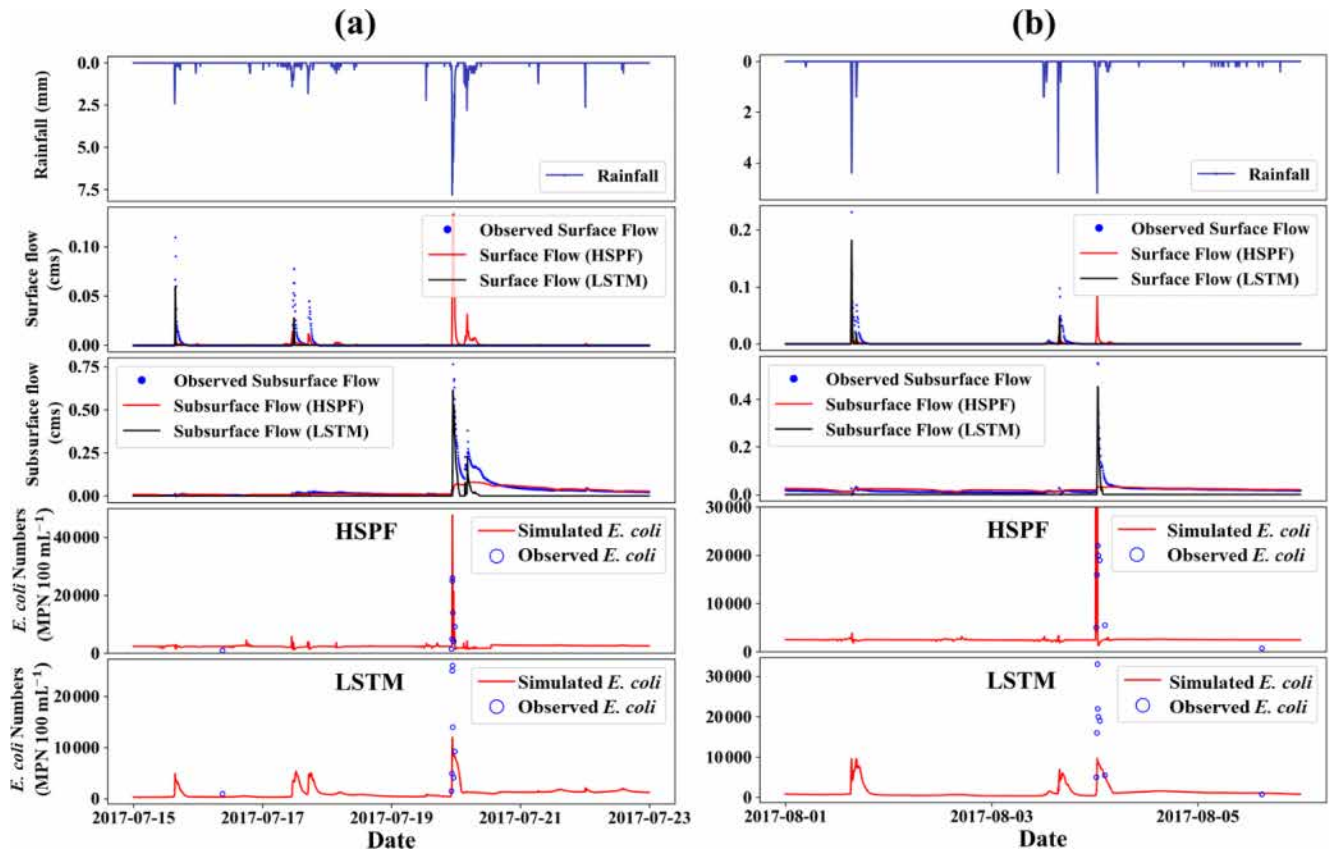


Figure 7. *E. coli* concentration simulated by HSPF and LSTM during 15–22 July (a) and 1–5 August 2017 (b). Both storm events were affiliated in the validation period.

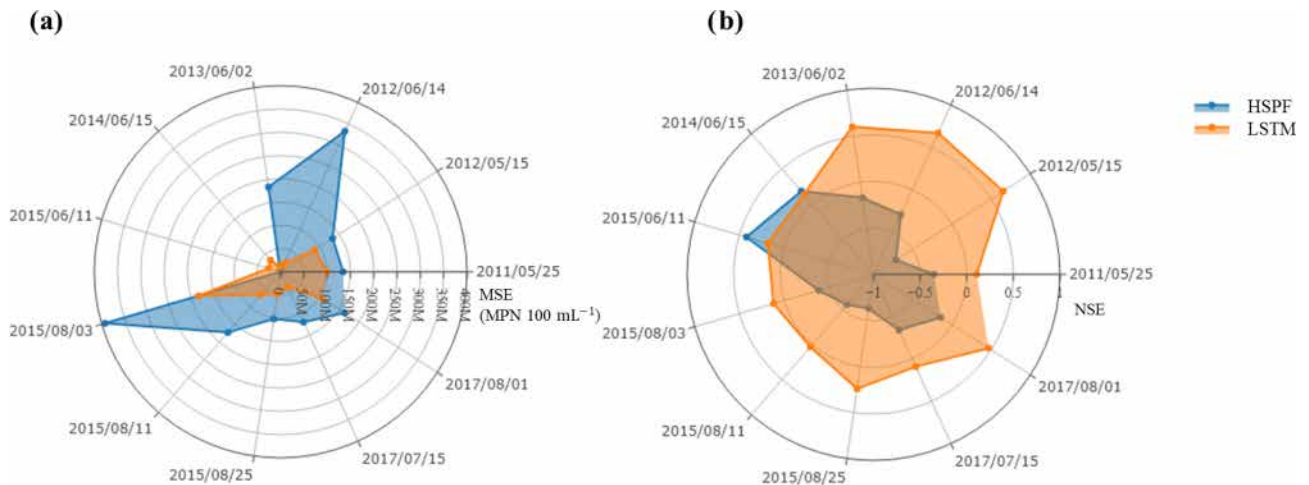


Figure 8. Comparison of the *E. coli* simulation during storm events: (a) MSE values and (b) NSE values.

ther sediment transport nor the attachment/detachment of *E. coli* on/from soil particles and suspended sediments. Several studies have been conducted on the monitoring and modeling of *E. coli* without considering sediment transport (Ahmadisharaf and Benham, 2020; Mishra et al., 2018). How-

ever, the need for its inclusion has been indicated elsewhere (Pandey and Soupir, 2013). To model sediment transport, additional data on suspended sediment concentration are required to build both the HSPF and deep-learning-based models. Therefore, this modeling can be further improved by col-

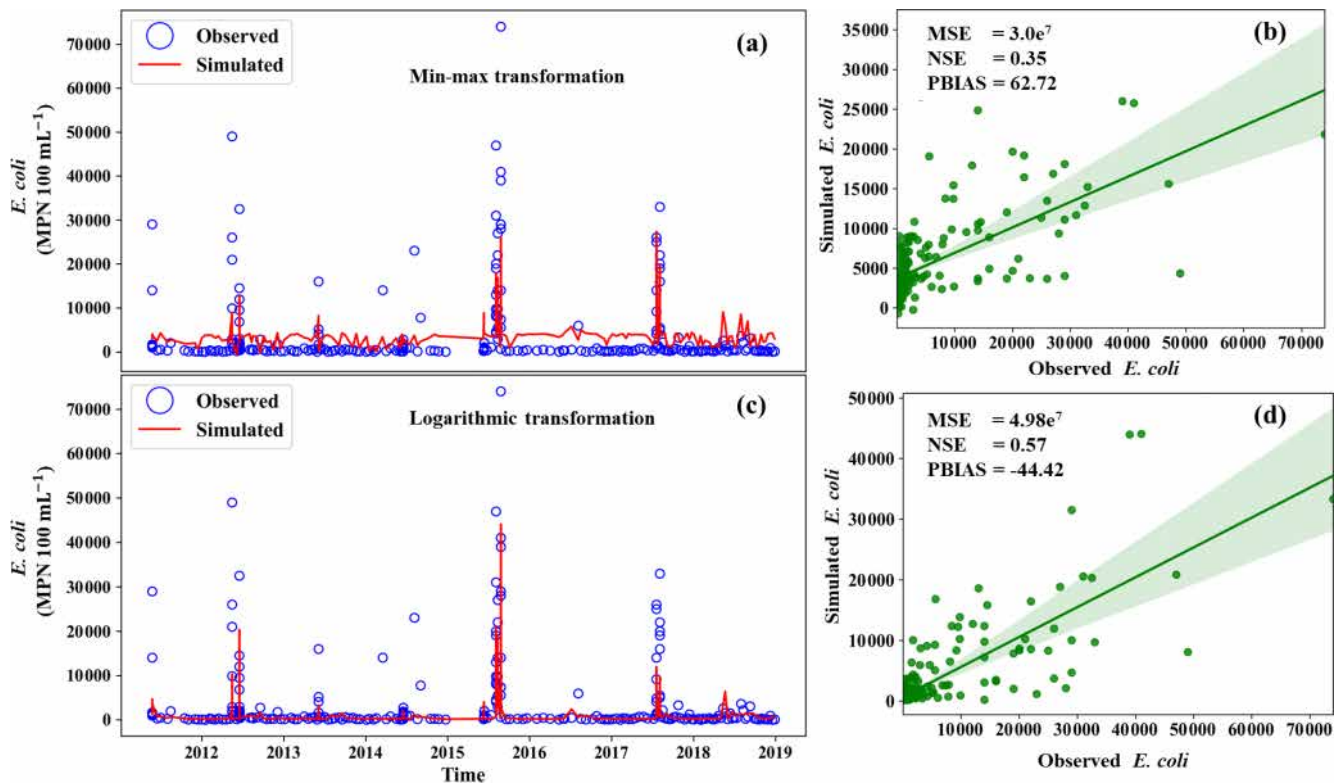


Figure 9. Comparison of *E. coli* concentration simulation with the transformation method: panels (a) and (c) indicate the *E. coli* simulation using min–max transformation and logarithmic transformation, respectively. Panels (b) and (d) indicate the scatter plot of *E. coli* using min–max transformation and logarithmic transformation, respectively.

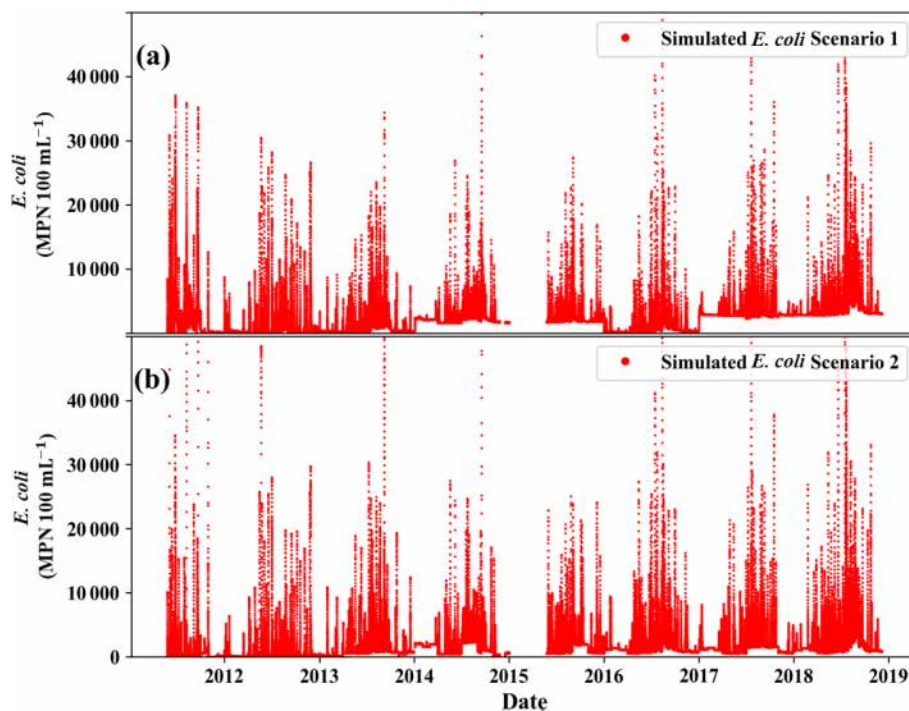


Figure 10. Simulated *E. coli* concentration with changes in *E. coli* sources with land-use change scenarios. Scenario 1 used land-use change and bacterial source information. Scenario 2 used the bacterial source by the fraction of each land use.

lecting sediment-related data and modeling sediment transport along with *E. coli* concentration.

The deep-learning-based approach can yield high model performance, but it has the limitation in terms of explainability and interpretability (Molnar, 2020). The neural networks are generally considered black boxes, and the question of interpreting them is still an open research problem (Mitchell, 2021; Tiddi, 2020). Several methods have been proposed to interpret the behavior of neural networks (Molnar et al., 2020). Explaining the output of neural networks can enhance the confidence of decision makers (Lipton, 2018). Therefore, we propose future research involving deep-learning models will benefit if the questions of interpretability and explainability are considered along with a model's prediction performance.

Deep-learning models are based upon the independent and identically distributed (IID) assumption, which means that the validation data are expected to have the same distribution as that of the training data (Kawaguchi et al., 2017). However, this is not a realistic assumption, and it is considered one of the challenges for researchers in machine learning (Bengio et al., 2021). Thus, in order to build regional or global hydrological models, the deep-learning model should be trained on catchment data from diverse catchments. Several researchers have adopted this approach to build regional models for streamflow prediction (Anderson and Radic, 2021; Kratzert et al., 2019; Xiang et al., 2021). However, a similar approach for building regional water quality models will be more challenging due to the scarcity of water quality data. We hope that the lessons from this study can be used as a guideline to train neural networks on regional water quality data.

4 Conclusions

In this study, we simulated the transport of bacteria in a head-water catchment of the northern Lao PDR at 6 min time steps. The main findings of this study are summarized as follows.

- Both the LSTM and HSPF models can accommodate land-use change and bacteria-source variation with time.
- The performance of the surface and subsurface flow simulation of LSTM was superior for both the calibration and validation steps when compared with the HSPF. The LSTM provided accurate surface and subsurface flow results by showing NSE values of 0.51 and 0.59, respectively, whereas the HSPF showed -0.7 and 0.55 of NSE.
- Our LSTM model showed better performance compared to HSPF for *E. coli* simulation. The NSE values of the HSPF and LSTM were -3.01 and 0.35, respectively. We found that the LSTM model can respond to changes in land use.

This study shows that deep-learning-based models are an efficient alternative to process-based models to simulate *E. coli* in a given catchment. Because LSTM can generate reasonable *E. coli* simulations, it could be applied to provide effective strategies for thwarting diseases that wreak havoc on human health. Therefore, a deep-learning approach can be useful in developing better water sustainability and management.

Code availability. The code is available on reasonable request from the corresponding author.

Data availability. The 6 min rainfall data (<https://doi.org/10.6096/MSEC.LAOS.5>, Silvera et al., 2015a), water-level data (<https://doi.org/10.6096/MSEC.LAOS.3>, Silvera et al., 2015b), land-use data (<https://doi.org/10.6096/MSEC.LAOS.7>, Silvera et al., 2015c), weather station data (<https://doi.org/10.6096/MSEC.LAOS.6>, Silvera et al., 2015d), *E. coli* data (<https://doi.org/10.23708/EWOYNK>, Ribolzi et al., 2021), soil map (<https://doi.org/10.23708/FFEDIR>, Chanhphengxay et al., 2021) and sub-catchment boundaries (<https://doi.org/10.23708/M8NJA0>, Boithias et al., 2021c) were collected from the M-TROPICS Critical Zone Observatory (<https://mtropics.obs-mip.fr/>, last access: 1 December 2021).

Supplement. The supplement related to this article is available online at: <https://doi.org/10.5194/hess-25-6185-2021-supplement>.

Author contributions. AA was responsible for conceptualization, data curation, methodology, visualization, writing, review and editing of the article. SB performed visualization, review and editing of the article. OR performed review and editing in addition to funding acquisition. NS performed data curation. BS was part of sampling and data acquisition. YP reviewed and edited the article. LB was part of funding acquisition, supervision, validation and review. KHC was part of conceptualization, funding acquisition, supervision, validation and review.

Competing interests. The authors declare that they have no conflict of interest.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Acknowledgements. This work was supported by the Korea Environment Industry and Technology Institute (KEITI) through the Aquatic Ecosystem Conservation Research Program, funded by the Korea Ministry of Environment (MOE) (RE202001319). The authors sincerely thank the Lao Department of Agricultural Land Management (DALaM) for its support, including granting the per-

mission for field access, and the M-TROPICS Critical Zone Observatory (<https://mtropics.obs-mip.fr/>, last access: 1 December 2021), which belongs to the French Research Infrastructure OZCAR (<http://www.ozcar-ri.org/>, last access: 1 December 2021), for data access. The authors also thank Campus France (PHC STAR 41510WH) for their financial support.

Financial support. This work was supported by the Korea Environment Industry and Technology Institute (KEITI) through the Aquatic Ecosystem Conservation Research Program, funded by the Korea Ministry of Environment (MOE) (RE202001319).

Review statement. This paper was edited by Thom Bogaard and reviewed by two anonymous referees.

References

- Abbasa, A., Baek, S., Kim M., Ligaray, M., Ribolzi, O., Silvera, N., Min, J.-H., Boithias, L., and Kyung, H. C.: Surface and sub-surface flow estimation at high temporal resolution using deep neural networks, *J. Hydrol.*, 590, 125370, <https://doi.org/10.1016/j.jhydrol.2020.125370>, 2020.
- Abimbola, O. P., Mittelstet, A. R., Messer, T. L., Berry, E. D., Bartelt-Hunt, S. L., and Hansen, S. P.: Predicting *Escherichia coli* loads in cascading dams with machine learning: An integration of hydrometeorology, animal density and grazing pattern, *Sci. Total Environ.*, 722, 137894, <https://doi.org/10.1016/j.scitotenv.2020.137894>, 2020.
- Abimbola, O., Mittelstet, A., Messer, T., Berry, E., and van Griensven, A.: Modeling and Prioritizing Interventions Using Pollution Hotspots for Reducing Nutrients, Atrazine and *E. coli* Concentrations in a Watershed, *Sustainability*, 13, 103, <https://doi.org/10.3390/su13010103>, 2021.
- Abadi, M., Barham, P., Chen, J., et al.: Kudlur, M.: Tensorflow: A system for large-scale machine learning. In 12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16), 265–283, Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation, usenix The advanced computing systems association, Berkeley, California, United States, 2016.
- Ackerman, D. and Weisberg, S. B.: Evaluating HSPF runoff and water quality predictions at multiple time and spatial scales, edited by: SBW a. K. Miller, Southern California coastal water research project biennial report, 2006, 3535 Harbor Blvd., Suite 110 Costa Mesa, CA 92626, USA, 293–303, 2005.
- Adomat, Y., Orzechowski, G. H., Pelger, M., Haas, R., Bartak, R., Nagy-Kovács, Z. Á., Appels, J., and Grischek, T.: New Methods for Microbiological Monitoring at Riverbank Filtration Sites, *Water*, 12, 584, <https://doi.org/10.3390/w12020584>, 2020.
- Ahmadisharaf, E. and Benham, B. L.: Risk-based decision making to evaluate pollutant reduction scenarios, *Sci. Total Environ.*, 702, 135022, <https://doi.org/10.1016/j.scitotenv.2019.135022>, 2020.
- Ahmed, S. I., Singh, A., Rudra, R., and Gharabaghi, B.: Comparison of CANWET and HSPF for water budget and water quality modeling in rural Ontario, *Water Qual. Res. J. Can.*, 49, 53–71, 2014.
- Anderson, S. and Radic, V.: Evaluation and interpretation of convolutional-recurrent networks for regional hydrological modelling, *Hydrol. Earth Syst. Sci. Discuss.* [preprint], <https://doi.org/10.5194/hess-2021-113>, in review, 2021.
- Banhatti, A. G. and Deka, P. C.: Effects of Data Pre-processing on the Prediction Accuracy of Artificial Neural Network Model in Hydrological Time Series, in: *Urban Hydrology, Watershed Management and Socio-Economic Aspects*, Springer, Heidelberg, Germany, 265–275, 2016.
- Bain, R. E., Wright, J. A., Christenson, E., and Bartram, J.: Rural: urban inequalities in post 2015 targets and indicators for drinking-water, *Sci. Total Environ.*, 490, 509–513, 2014.
- Bengio, Y., Lecun, Y., and Hinton, G.: Deep learning for AI, *Commun. ACM*, 64, 58–65, 2021.
- Benham, B., Yagow, G., Barham, B., Zeckoski, R., and Dillaha, T.: Total Maximum Daily Load Development: Mill Creek bacteria (*E. coli*) impairment, Page County, Virginia, Richmond, VA, USA, Virginia Department of Environmental Quality, 2005.
- Bicknell, B. R., Imhoff, J. C., Kittle Jr., J. L., Donigan Jr., A. S., and Johanson, R. C.: Hydrological simulation program – FORTRAN user’s manual for version 11, Environmental Protection Agency Report No. EPA/600/R-97/080, US Environmental Protection Agency, Athens, GA, USA, 1997.
- Boithias, L., Choisy, M., Souliyaseng, N., Jourden, M., Quet, F., Buisson, Y., Thammahacksa, C., Silvera, N., Latsachack, K., Sengtaheuanghoung, O., Pierret, A., Rochelle-Newall, E., Becerra, S., and Ribolzi, O.: Hydrological regime and water shortage as drivers of the seasonal incidence of diarrheal diseases in a tropical montane environment, *PLOS Neglect. Trop. D.*, 10, e0005195. <https://doi.org/10.1371/journal.pntd.0005195>, 2016.
- Boithias, L., Auda, Y., Audry, S., Bricquet, J. P., Chanhphengxay, A., Chaplot, V., de Rouw, A., Henry des Tureaux, T., Huon, S., and Janeau, J. I.: The Multiscale TROPICAL CatchmentS critical zone observatory M-TROPICS dataset II: land use, hydrology and sediment production monitoring in Houay Pano, northern Lao PDR, *Hydrol. Proc.*, 35, e14126, <https://doi.org/10.1002/hyp.14126>, 2021a.
- Boithias, L., Ribolzi, O., Lacombe, G., Thammahacksa, C., Silvera, N., Latsachack, K., Soulieth, B., Viguier, M., Auda, Y., and Robert, E.: Quantifying the effect of overland flow on *Escherichia coli* pulses during floods: use of a tracer-based approach in an erosion-prone tropical catchment, *J. Hydrol.*, 594, 125935, <https://doi.org/10.1016/j.jhydrol.2020.125935>, 2021b.
- Boithias, L., Ribolzi, O., Phachomphon, K., Phommasack, T., Valentin, C., and Sipaseuth, N.: Sub-catchments boundaries of the Houay Pano catchment, northern Lao PDR [Data set], <https://doi.org/10.23708/M8NJA0>, 2021c.
- Cause, J., Billen, G., Garnier, J., Henri-des-Tureaux, T., Olassa, X., Thammahacksa, C., Latsachack, K. O., Soulieth, B., Sengtaheuanghoung, O., Rochelle-Newall, E., and Ribolzi, O.: Field and modelling studies of *Escherichia coli* loads in tropical streams of montane agroecosystems, *J. Hydro.-Environ. Res.*, 9, 496–507, <https://doi.org/10.1016/j.jher.2015.03.003>, 2015.
- Chanhphengxay, A., Phommasack, T., and Valentin, C.: Soil map of the Houay Pano catchment, northern Lao PDR (1998) [Data set], <https://doi.org/10.23708/FFEDIR>, 2021.

- Chen, H. J. and Chang, H.: Response of discharge, TSS, and *E. coli* to rainfall events in urban, suburban, and rural watersheds, *Environmental Science: Processes & Impacts*, 16, 2313–2324, 2014.
- Chen, K., Chen, H., Zhou, C., Huang, Y., Qi, X., Shen, R., Liu, F., Zuo, M., Zoua, X., Wang, J., Zhang, Y., Chen, D., Chen, X., Deng, Y., and Renc, H.: Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data, *Water Res.*, 171, 115454, <https://doi.org/10.1016/j.watres.2019.115454>, 2020.
- Chin, D. A., Sakura-Lemessy, D., Bosch, D. D., and Gay, P. A.: Watershed-scale fate and transport of bacteria, *T. ASABE*, 52, 145–154, <https://doi.org/10.13031/2013.25955>, 2009.
- Cho, K. H., Pachepsky, Y. A., Kim, J. H., Guber, A. K., Shelton, D. R., and Rowland, R.: Release of *Escherichia coli* from the bottom sediment in a first-order creek: Experiment and reach-specific modeling, *J. Hydrol.*, 391, 322–332, <https://doi.org/10.1016/j.jhydrol.2010.07.033>, 2010.
- Cho, K. H., Pachepsky, Y. A., Oliver, D. M., Muirhead, R. W., Park, Y., Quilliam, R. S., and Shelton, D. R.: Modeling fate and transport of fecally-derived microorganisms at the watershed scale: state of the science and future opportunities, *Water Res.*, 100, 38–56, <https://doi.org/10.1016/j.watres.2016.04.064>, 2016.
- Chuang, C. C., Wang, C. M., and Li, C. W.: Weighted linear regression for symbolic interval-values data with outliers, in: 2010 5th IEEE Conference on Industrial Electronics and Applications, IEEE, 2238–2242, 15–17 June 2010, Taichung, Taiwan, <https://doi.org/10.1109/ICIEA.2010.5515157>, 2010.
- Clevert, D. A., Unterthiner, T., and Hochreiter, S.: Fast and accurate deep network learning by exponential linear units (elus), *arXiv [preprint]*, [arXiv:1511.07289](https://arxiv.org/abs/1511.07289), 2015.
- Chollet, F.: Deep learning with Python, Simon and Schuster, Manning Publications Co, 20 Baldwin Road, P.O. Box 761, Shelter Island, NY 11964, USA, ISBN 9781617294433, 2018.
- Dosovitskiy, A. and Djolonga, J.: You Only Train Once: Loss-Conditional Training of Deep Networks, in: International Conference on Learning Representations, available at: <https://openreview.net/pdf?id=HyxY6JHKwr>, last access: September 2019.
- Dong, Q., Lin, Y., Bi, J., and Yuan, H.: An Integrated Deep Neural Network Approach for Large-Scale Water Quality Time Series Prediction, in: 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC), 6–9 October 2019, <https://doi.org/10.1109/SMC.2019.8914404>, IEEE, Bari, Italy, 3537–3542, 2019.
- Ferguson, C. M., Croke, B. F., Beatson, P. J., Ashbolt, N. J., and Deere, D. A.: Development of a process-based model to predict pathogen budgets for the Sydney drinking water catchment, *J. Water Health*, 5, 187–208, 2007.
- Fonseca, A., Botelho, C., Boaventura, R. A., and Vilar, V. J.: Integrated hydrological and water quality model for river management: a case study on Lena River, *Sci. Total Environ.*, 485, 474–489, 2014.
- Frolich, L., Vaizel-Ohayon, D., and Fishbain, B.: Prediction of Bacterial Contamination Outbursts in Water Wells through Sparse Coding, *Sci. Rep.*, 7, 1–11, <https://doi.org/10.1038/s41598-017-00830-4>, 2017.
- Fujioka, R. S., Solo-Gabriele, H. M., Byappanahalli, M. N., and Kirs, M. US recreational water quality criteria: a vision for the future, *Int. J. Env. Res. Pub. He.*, 12, 7752–7776, <https://doi.org/10.3390/ijerph120707752>, 2015.
- Gaillardet, J., Braud, I., Hankard, F., et al.: OZCAR: The French network of critical zone observatories, *Vadose Zone J.*, 17, 1–24, <https://doi.org/10.2136/vzj2018.04.0067>, 2018.
- Gassman, P. W., Reyes, M. R., Green, C. H., and Arnold, J. G.: The soil and water assessment tool: historical development, applications, and future research directions, *T. ASABE*, 50, 1211–1250, <https://doi.org/10.13031/2013.23637>, 2007.
- Goodfellow, I., Bengio, Y., and Courville, A.: Deep learning: MIT press, Cambridge, Massachusetts, USA, 2016.
- Gupta, H. V., Sorooshian, S., and Yapo, P. O.: Status of automatic calibration for hydrologic models: Comparison with multilevel expert calibration, *J. Hydrol. Eng.*, 4, 135–143, [https://doi.org/10.1061/\(ASCE\)1084-0699\(1999\)4:2\(135\)](https://doi.org/10.1061/(ASCE)1084-0699(1999)4:2(135)), 1999.
- Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, *J. Hydrol.*, 377, 80–91, <https://doi.org/10.1016/j.jhydrol.2009.08.003>, 2009.
- Heaphy, R. T., Burke, M. P., and Love, J. T.: Conversion of HSPF Legacy Model to a Platform-Independent, Open-Source Language. AGUFM, 2015, H13C-1529, American Geophysical Union, Fall Meeting 2015, H13C-1529, December 2015.
- Hinton, G. E., Osindero, S., and Teh, Y. W.: A fast learning algorithm for deep belief nets, *Neural Comput.*, 18, 1527–1554, <https://doi.org/10.1162/neco.2006.18.7.1527>, 2006.
- Hochreiter, S. and Schmidhuber, J.: Long short-term memory, *Neural Comput.*, 9, 1735–1780, 1997.
- Iqbal, M. S., Islam, M. M., and Hofstra, N.: The impact of socio-economic development and climate change on *E. coli* loads and concentrations in Kabul River, Pakistan, *Sci. Total Environ.*, 650, 1935–1943, <https://doi.org/10.1016/j.scitotenv.2018.09.347>, 2019.
- Isikdogan, F., Bovik, A. C., and Passalacqua, P.: Surface water mapping by deep learning, *IEEE J. Sel. Top. Appl.*, 10, 4909–4918, 2017.
- Kawaguchi, K., Kaelbling, L. P., and Bengio, Y.: Generalization in deep learning, *arXiv [preprint]*, [arXiv:1710.05468](https://arxiv.org/abs/1710.05468), 2017.
- Kim, M., Boithias, L., Cho, K. H., Silvera, N., Thammahacksa, C., Latsachack, K., Rochelle-Newall, E., Sengtaheuanghoung, O., Pierret, A., Pachepsky, Y. A., and Ribolzi, O.: Hydrological modeling of fecal indicator bacteria in a tropical mountain catchment, *Water Res.*, 119, 102–113, <https://doi.org/10.1016/j.watres.2017.04.038>, 2017.
- Kim, M., Boithias, L., Cho, K. H., Sengtaheuanghoung, O., and Ribolzi, O.: Modeling the Impact of Land Use Change on Basin-scale Transfer of Fecal Indicator Bacteria: SWAT Model Performance, *J. Environ. Qual.*, 47, 1115–1122, 2018.
- Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., and Nearing, G.: Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets, *Hydrol. Earth Syst. Sci.*, 23, 5089–5110, <https://doi.org/10.5194/hess-23-5089-2019>, 2019.
- Lee, D. H., Kim, J. H., Park, M.-H., Stenstrom, M. K., and Kang, J.-H.: Automatic calibration and improvements on an instream

- chlorophyll a simulation in the HSPF model, *Ecol. Modell.*, 415, 108835, <https://doi.org/10.1016/j.ecolmodel.2019.108835>, 2020.
- Lin, F., Chen, X., and Yao, H.: Evaluating the use of Nash-Sutcliffe efficiency coefficient in goodness-of-fit measures for daily runoff simulation with SWAT, *J. Hydrol. Eng.*, 22, 05017023, [https://doi.org/10.1061/\(ASCE\)HE.1943-5584.0001580](https://doi.org/10.1061/(ASCE)HE.1943-5584.0001580), 2017.
- Lipton, Z. C.: The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery, *Queue*, 16, 31–57, 2018.
- Mazzocchi, F.: Could Big Data be the end of theory in science? A few remarks on the epistemology of data-driven science, *EMBO reports*, 16, 1250–1255, <https://doi.org/10.15252/embr.201541001>, 2015.
- Mishra, A., Ahmadisharaf, E., Benham, B. L., Wolfe, M. L., Leman, S. C., Gallagher, D. L., Reckhow, K. H., and Smith, E. P.: Generalized likelihood uncertainty estimation and Markov chain Monte Carlo simulation to prioritize TMDL pollutant allocations, *J. Hydrol. Eng.*, 23, 05018025, [https://doi.org/10.1061/\(ASCE\)HE.1943-5584.0001720](https://doi.org/10.1061/(ASCE)HE.1943-5584.0001720), 2018.
- Mitchell, M.: Why AI is harder than we think, arXiv [preprint], arXiv:2104.12871, 2021.
- Molnar, C.: Interpretable machine learning, Lulu.com, available at: <https://christophm.github.io/interpretable-ml-book/> (last access: 1 December 2021), 2020.
- Molnar, C., Casalicchio, G., and Bischl, B.: Interpretable machine learning – a brief history, state-of-the-art and challenges, Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 417–431, 14–18 September 2020, Ghent, Belgium, https://doi.org/10.1007/978-3-030-65965-3_28, 2020.
- Morris, M. D.: Factorial sampling plans for preliminary computational experiments, *Technometrics*, 33, 161–174, 1991.
- Meshesha, T. W., Wang, J., and Melaku, N. D.: A modified hydrological model for assessing effect of pH on fate and transport of *Escherichia coli* in the Athabasca River basin, *J. Hydrol.*, 582, 124513, <https://doi.org/10.1016/j.jhydrol.2019.124513>, 2020.
- Moriasi, D. N., Gitau, M. W., Pai, N., and Daggupati, P.: Hydrologic and water quality models: Performance measures and evaluation criteria, *T. ASABE*, 58, 1763–1785, 2015.
- Muirhead, R. W. and Meenken, E. D.: Variability of *Escherichia coli* Concentrations in Rivers during Base-Flow Conditions in New Zealand, *J. Environ. Qual.*, 47, 967–973, <https://doi.org/10.2134/jeq2017.11.0458>, 2018.
- Nakhle, P., Ribolzi, O., Boithias, L., Rattanavong, S., Auda, Y., Sayavong, S., Zimmermann, R., Soullieuth, B., Pando, A., and Thammahacksa, C.: Effects of hydrological regime and land use on in-stream *Escherichia coli* concentration in the Mekong basin, *Lao PDR, Sci. Rep.*, 11, 1–17, 2021.
- Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models part I – A discussion of principles, *J. Hydrol.*, 10, 282–290, 1970.
- Nash, S. G.: Newton-type minimization via the Lanczos method, *SIAM J. Numer. Anal.*, 21, 770–788, 1984.
- Nguyen, H. T. M., Le, Q. T. P., Garnier, J., Janeau, J. L., and Rochelle-Newall, E.: Seasonal variability of faecal indicator bacteria numbers and die-off rates in the Red River basin, North Viet Nam, *Sci. Rep.*, 6, 1–12, <https://doi.org/10.1038/srep21644>, 2016.
- Nair, V. and Hinton, G. E.: Rectified linear units improve restricted boltzmann machines, in: *ICML, Proceedings of the 27th International Conference on Machine Learning*, 807–814, June, 2010.
- Neitsch, S. L., Arnold, J. G., Kiniry, J. R., and Williams, J. R.: Soil and water assessment tool theoretical documentation version 2009, Texas Water Resources Institute, Texas, USA, 2011.
- Odonkor, S. T. and Ampofo, J. K.: *Escherichia coli* as an indicator of bacteriological quality of water: an overview, *Microbiology research*, 4, e2–e2, <https://doi.org/10.4081/mr.2013.e2>, 2013.
- Palmateer, G., McLean, D., Kutas, W. L., and Meissner, S. M.: Suspended particulate/bacterial interaction in agricultural drains, *SS RAO*, 1–40, CRC Press Inc., Florida, 1993.
- Pachepsky, Y. and Shelton, D.: *Escherichia coli* and fecal coliforms in freshwater and estuarine sediments, *Crit. Rev. Env. Sci. Tec.*, 41, 1067–1110, 2011.
- Pachepsky, Y. A., Blaustein, R. A., Whelan, G., and Shelton, D. R.: Comparing temperature effects on *Escherichia coli*, *Salmonella*, and *Enterococcus* survival in surface waters, *Lett. Appl. Microbiol.*, 59, 278–283, <https://doi.org/10.1111/lam.12272>, 2014.
- Pachepsky, Y., Stocker, M., Saldaña, M. O., and Shelton, D.: Enrichment of stream water with fecal indicator organisms during baseflow periods, *Environ. Monit. Assess.*, 189, 51, <https://doi.org/10.1007/s10661-016-5763-8>, 2017.
- Pachepsky, Y. A., Allende, A., Boithias, L., Cho, K., Jamieson, R., Hofstra, N., and Molina, M.: Microbial water quality: monitoring and modeling, *J. Environ. Qual.*, 47, 931–938, <https://doi.org/10.2134/jeq2018.07.0277>, 2018.
- Pandey, P. K. and Soupir, M. L.: Assessing the impacts of *E. coli* laden streambed sediment on *E. coli* loads over a range of flows and sediment characteristics, *J. Am. Water Resour. As.*, 49, 1261–1269, <https://doi.org/10.1111/jawr.12079>, 2013.
- Park, Y., Kim, M., Pachepsky, Y., Choi, S. H., Cho, J. G., Jeon, J., and Cho, K. H.: Development of a nowcasting system using machine learning approaches to predict fecal contamination levels at recreational beaches in Korea, *J. Environ. Qual.*, 47, 1094–1102, <https://doi.org/10.2134/jeq2017.11.0425>, 2018.
- Patin, J., Mouche, E., Ribolzi, O., Sengtahevanguong, O., Latsachak, K., Soullieuth, B., Chaplot, V., and Valentin, C.: Effect of land use on interrill erosion in a montane catchment of Northern Laos: An analysis based on a pluri-annual runoff and soil loss database, *J. Hydrol.*, 563, 480–494, 2018.
- Peterson, K. T., Sagan, V., and Sloan, J. J.: Deep learning-based water quality estimation and anomaly detection using Landsat-8/Sentinel-2 virtual constellation and cloud computing, *Gisci. Remote Sens.*, 57, 510–525, 2020.
- Pool, S., Vis, M., and Seibert, J.: Evaluating model performance: towards a non-parametric variant of the Kling-Gupta efficiency, *Hydrol. Sci. J.*, 63, 1941–1953, 2018.
- Pyo, J., Park, L. J., Pachepsky, Y., Baek, S. S., Kim, K., and Cho, K. H.: Using convolutional neural network for predicting cyanobacteria concentrations in river water, *Water Res.*, 186, 116349, <https://doi.org/10.1016/j.watres.2020.116349>, 2020.
- Read, J. S., Jia, X., Willard, J., Appling, A. P., Zwart, J. A., Oliver, S. K., Karpatne, A., Hansen, G. J. A., Hanson, P. C., Watkins, W., Steinbach, M., and Kumar, V.: Process-guided deep learning predictions of lake water temperature, *Water Resour. Res.*, 55, 9173–9190, <https://doi.org/10.1029/2019WR024922>, 2019.
- Rochelle-Newall, E., Nguyen, T. M. H., Le, T. P. Q., Sengtahevanguong, O., and Ribolzi, O.: A short review of fe-

- cal indicator bacteria in tropical aquatic ecosystems: knowledge gaps and future directions, *Front. Microbiol.*, 6, 308, <https://doi.org/10.3389/fmicb.2015.00308>, 2015.
- Rochelle-Newall, E. J., Ribolzi, O., Viguier, M., Thammahacksa, C., Silvera, N., Latsachack, K., Dinh, R. P., Naporn, P., Sy, H. T., and Soullileuth, B.: Effect of land use and hydrological processes on *Escherichia coli* concentrations in streams of tropical, humid headwater catchments, *Sci. Rep.*, 6, 1–12, 2016.
- Ribolzi, O., Evrard, O., Huon, S., Rochelle-Newall, E., Henri-des-Tureaux, T., Silvera, N., Thammahacksa, C., and Sengtaheuanghoung, O.: Use of fallout radionuclides (^7Be , ^{210}Pb) to estimate resuspension of *Escherichia coli* from streambed sediments during floods in a tropical montane catchment, *Environ. Sci. Pollut. R.*, 23, 3427–3435, 2016.
- Ribolzi, O., Evrard, O., Huon, S., De Rouw, A., Silvera, N., Latsachack, K. O., Soullileuth, B., Lefèvre, I., Pierret, A., and Lacombe, G.: From shifting cultivation to teak plantation: effect on overland flow and sediment yield in a montane tropical catchment, *Sci. Rep.*, 7, 1–12, 2017.
- Ribolzi, O., Lacombe, G., Pierret, A., Robain, H., Sounyafong, P., De Rouw, A., Soullileuth, B., Mouche, E., Huon, S., and Silvera, N.: Interacting land use and soil surface dynamics control groundwater outflow in a montane catchment of the lower Mekong basin, *Agr. Ecosyst. Environ.*, 268, 90–102, 2018.
- Ribolzi, O., Boithias, L., Thammahacksa, C., Rochelle-Newall, E., Pando-Bahuon, A., Silvera, N., Sengtaheuanghoung, O., Sipaseuth, N., and Pierret, A.: *Escherichia coli* concentrations and physico-chemical measurements (2011–2021) at the outlet of the Houay Pano catchment, northern Lao PDR [Data set], <https://doi.org/10.23708/EWOYNK>, 2021.
- Rumelhart, D. E., Hinton, G. E., and Williams, R. J.: Learning representations by back-propagating errors, *Nature*, 323, 533–536, <https://doi.org/10.1038/323533a0>, 1986.
- Seong, C. H., Benham, B. L., Hall, K. M., and Kline, K.: Comparison of alternative methods to simulate bacteria concentrations with HSPF under low-flow conditions, *Appl. Eng. Agric.*, 29, 917–931, <https://doi.org/10.13031/aea.29.10203>, 2013.
- Silvera, N., Ribolzi, O., Boithias, L., Rochelle-Newall, E., Riotte, J., Audry, S., Sipaseuth, N., Valentin, C., Janeau, J. L., Bricquet, J. P., Sengtaheuanghoung, O., Auda, Y., Chaplot, V., de Rouw, A., Henry-Des-Tureaux, T., Huon, S., Latsachack, K., Maeght, J. L., Pando, A., Pierret, A., Robain, H., Sayavong, S., Soullileuth, B., Souliyavongsa, X., Sounyafong, P., Thammahacksa, C., A., Viguier, M., Khampaseuth, X., Bourdon, E., Chanhphengxay, A., Le Troquer, Y., Lestrelin, G., Marchand, P., Moreau, P., Phachomphon, K., Phantahvong, K., Tasaketh, S., Thieboux, J., Vi-giak, O., and Noble, A.: 6 min rainfall data, Houay Pano, Laos [Data set], <https://doi.org/10.6096/msec.laos.5>, 2015a.
- Silvera, N., Ribolzi, O., Boithias, L., Rochelle-Newall, E., Riotte, J., Audry, S., Sipaseuth, N., Valentin, C., Janeau, J. L., Bricquet, J. P., Sengtaheuanghoung, O., Auda, Y., Chaplot, V., de Rouw, A., Henry-Des-Tureaux, T., Huon, S., Latsachack, K., Maeght, J. L., Pando, A., Pierret, A., Robain, H., Sayavong, S., Soullileuth, B., Souliyavongsa, X., Sounyafong, P., Thammahacksa, C., A., Viguier, M., Khampaseuth, X., Bourdon, E., Chanhphengxay, A., Le Troquer, Y., Lestrelin, G., Marchand, P., Moreau, P., Phachomphon, K., Phantahvong, K., Tasaketh, S., Thieboux, J., Vi-giak, O., and Noble, A.: Hydrological data, Houay Pano, Laos [Data set], <https://doi.org/10.6096/msec.laos.3>, 2015b.
- Silvera, N., Ribolzi, O., Boithias, L., Rochelle-Newall, E., Riotte, J., Audry, S., Sipaseuth, N., Valentin, C., Janeau, J. L., Bricquet, J. P., Sengtaheuanghoung, O., Auda, Y., Chaplot, V., de Rouw, A., Henry-Des-Tureaux, T., Huon, S., Latsachack, K., Maeght, J. L., Pando, A., Pierret, A., Robain, H., Sayavong, S., Soullileuth, B., Souliyavongsa, X., Sounyafong, P., Thammahacksa, C., A., Viguier, M., Khampaseuth, X., Bourdon, E., Chanhphengxay, A., Le Troquer, Y., Lestrelin, G., Marchand, P., Moreau, P., Phachomphon, K., Phantahvong, K., Tasaketh, S., Thieboux, J., Vi-giak, O., and Noble, A.: Land use data, Houay Pano, Laos [Data set], <https://doi.org/10.6096/msec.laos.7>, 2015c.
- Silvera, N., Ribolzi, O., Boithias, L., Rochelle-Newall, E., Riotte, J., Audry, S., Sipaseuth, N., Valentin, C., Janeau, J. L., Bricquet, J. P., Sengtaheuanghoung, O., Auda, Y., Chaplot, V., de Rouw, A., Henry-Des-Tureaux, T., Huon, S., Latsachack, K., Maeght, J. L., Pando, A., Pierret, A., Robain, H., Sayavong, S., Soullileuth, B., Souliyavongsa, X., Sounyafong, P., Thammahacksa, C., A., Viguier, M., Khampaseuth, X., Bourdon, E., Chanhphengxay, A., Le Troquer, Y., Lestrelin, G., Marchand, P., Moreau, P., Phachomphon, K., Phantahvong, K., Tasaketh, S., Thieboux, J., Vi-giak, O., and Noble, A.: Weather station data, Houay Pano, Laos [Data set], <https://doi.org/10.6096/msec.laos.6>, 2015d.
- Singh, A. and Kingsbury, N.: Dual-tree wavelet scattering network with parametric log transformation for object classification, in: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2622–2626, IEEE, <https://doi.org/10.1109/ICASSP.2017.7952631>, March 2017.
- Solanki, A., Agrawal, H., and Khare, K.: Predictive analysis of water quality parameters using deep learning, *Int. J. Comp. Appl.*, 125, 0975–8887, <https://doi.org/10.5120/ijca2015905874>, 2015.
- Song, L., Boithias, L., Sengtaheuanghoung, O., Oeurng, C., Valentin, C., Souksavath, B., Sounyafong, P., de Rouw, A., Soullileuth, B., Silvera, N., Lattanavongkot, B., Pierret, A., and Ribolzi, O.: Understory Limits Surface Runoff and Soil Loss in Teak Tree Plantations of Northern Lao PDR, *Water*, 12, 2327, <https://doi.org/10.3390/w12092327>, 2020.
- Sowah, R. A., Bradshaw, K., Snyder, B., Spidle, D., and Molina, M.: Evaluation of the soil and water assessment tool (SWAT) for simulating *E. coli* concentrations at the watershed-scale, *Sci. Total Environ.*, 746, 140669, <https://doi.org/10.1016/j.scitotenv.2020.140669>, 2020. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting, *J. Mach. Learn. Res.*, 15, 1929–1958, 2014.
- Sze, V., Chen, Y. H., Yang, T. J., and Emer, J. S.: Efficient Processing of Deep Neural Networks: A Tutorial and Survey, <https://doi.org/10.1109/JPROC.2017.2761740>, Proceedings of the IEEE, 2295–2329, 2017.
- Thupaki, P., Phanikumar, M. S., Schwab, D. J., Nevers, M. B., and Whitman, R. L.: Evaluating the role of sediment-bacteria interactions on *Escherichia coli* concentrations at beaches in southern Lake Michigan, *J. Geophys. Res.-Oceans*, 118, 7049–7065, <https://doi.org/10.1002/2013JC008919>, 2013.
- Troeger, C., Forouzanfar, M., Rao, P. C., et al.: Estimates of global, regional, and national morbidity, mortality, and aetiologies of diarrhoeal diseases: a systematic analysis for the Global Burden of Disease Study 2015, *Lancet. Infect. Dis.*, 17, 909–948, [https://doi.org/10.1016/S1473-3099\(17\)30276-1](https://doi.org/10.1016/S1473-3099(17)30276-1), 2017.

- Tiddi, I.: Directions for explainable knowledge-enabled systems, Knowledge Graphs for eXplainable Artificial Intelligence: Foundations, Applications and Challenges, 47, 245, ISBN 978-1-64368-080-4, 245–261, <https://doi.org/10.3233/SSW200022>, 2020.
- Van Rossum, G.: Python programming language, in: USENIX annual technical conference, Vol. 41, p. 36, Santa Clara, CA, USA, June 2007.
- Virtanen, P., Gommers, R., Oliphant, T. E., et al.: SciPy 1.0: fundamental algorithms for scientific computing in Python, Nat. Methods, 17, 261–272, <https://doi.org/10.1038/s41592-019-0686-2>, 2020
- Wang, X., Zhang, F., and Ding, J.: Evaluation of water quality based on a machine learning algorithm and water quality index for the Ebinur Lake Watershed, China, Sci. Rep., 7, 1–18, <https://doi.org/10.1038/s41598-017-12853-y>, 2017
- Whitehead, P. G., Leckie, H., Rankinen, K., Butterfield, D., Futter, M., and Bussi, G.: An INCA model for pathogens in rivers and catchments: Model structure, sensitivity analysis and application to the River Thames catchment, UK, Sci. Total Environ., 572, 1601–1610, 2016.
- Xiang, Z., Yan, J., and Demir, I.: A rainfall-runoff model with LSTM-based sequence-to-sequence learning, Water Resour. Res., 56, e2019WR025326, <https://doi.org/10.1029/2019WR025326>, 2020.
- Xiang, Z., Demir, I., Mantilla, R., and Krajewski Witold, F.: A Regional Semi-Distributed Streamflow Model Using Deep Learning, <https://doi.org/10.31223/X5GW3V>, 2011.
- Van der Leeuw, S. E.: Why model?, Cybernet. Syst., 35, 117–128, <https://doi.org/10.1080/01969720490426803>, 2004.
- Yagow, G., Dillaha, T., Mostaghimi, S., Brannan, K., Heatwole, C., and Wolfe, M. L.: TMDL modeling of fecal coliform bacteria with HSPF, in: 2001 ASAE Annual Meeting, p. 1, American Society of Agricultural and Biological Engineers, 2950 Niles Road, St. Joseph, MI 49085, Copyright © 2021 American Society of Agricultural and Biological Engineers, 1998.
- Zheng, A. and Casari, A.: Feature engineering for machine learning: principles and techniques for data scientists, O'Reilly Media, Inc., 1005 Gravenstein Highway North Sebastopol, CA 95472, USA, 2018.