Supplementary Information for

In-stream *Escherichia Coli* Modeling Using high-temporal-resolution data with deep learning and process-based models

Ather Abbas[1], *Sangsoo Baek*[1], *Norbert Silvera*[2], *Bounsamay Soulileuth*[3], *Yakov Pachepsky*[4] *Olivier Ribolzi*[5], *Laurie Boithias*[5†], *Kyung Hwa Cho*[1†]

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

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

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

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

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

[†] *Co-corresponding authors: Kyung Hwa Cho (*khcho@unist.ac.kr*), Laurie Boithias (*laurie.boithias@get.omp.eu*)*

A manuscript for

Hydrology and Earth System Sciences

Text S1. Study area and land use information

The 0.6 km^2 catchment is part of the 800000 km^2 Mekong River basin. The catchment is located at an altitude of 435–716 m (Fig. 1) with a slope gradient of 1 %–135 % (mean = 52 %). The closest village is Lak Sip, located downstream of station S4 (Fig. 1), which has 484 inhabitants (Census of 2015).

The study area can be characterized as subhumid with a monsoon season. The dry season stretches from November to May, whereas the wet season spans from June to October. The annual mean temperature is 23.4 °C, and the annual mean precipitation is 1,366 mm from 2001 to 2019 (Boithias et al., 2021). However, during our study period (i.e., from 2011 to 2018), the mean annual precipitation was 1450 mm. Approximately 71 % of rainfall occurs during the wet season. The subsurface geology predominantly consists of Permian to Upper Carboniferous argillites, siltstones, and fine-grained sandstones. The soils in the study area can be classified as Entisol, Ultisol, and Alfisol, comprising 20 %, 30 %, and 50 %, respectively.

Detailed land-cover surveys and mapping were conducted each year from 2011 to 2018 within the catchment area. The annual areal percentages of fallow, teak trees, annual crops, and forest were calculated using a geospatial information processing software QGIS version 2.6 (QGIS Development Team, 2016) and denoted “Fallow,” “Teak,” “Annual crop,” and “Forest,” for modeling purposes, respectively. The land-use change for each type of land use is shown in the form of time series in Fig. S1. The area has recently

undergone an increase in teak tree plantations, especially from 2006 to 2013 (Ribolzi et al., 2017). The fallow land use also increased at the expense of annual crops from 2012 to 2016.

Text S2. Electrical-conductivity-based hydrograph separation

We used a tracer-based approach (Collins and Neal, 1998) to separate storm hydrographs into "event water" (infiltration-excess overland flow) and "pre-event water" (groundwater pre-stored in the catchment area). This approach relies on a simple mixing model with two reservoirs and the electrical conductivity of water as a tracer, and it had been previously tested in the study catchment by Ribolzi et al., 2018). The tracer-based approach is described by the following equations:

$$Q = Q_{OF} + Q_{GW}, \quad (1)$$

$$Q \times EC = Q_{OF} \times EC_{OF} + Q_{GW} \times EC_{GW}, \quad (2)$$

where Q is the instantaneous stream water discharge rate at the catchment outlet ($m^3 s^{-1}$); Q_{OF} is the instantaneous discharge of overland flow—surface flow ($m^3 s^{-1}$); Q_{GW} is the instantaneous discharge of groundwater—subsurface flow ($m^3 s^{-1}$); EC is the instantaneous electrical conductivity measured in the stream (µS cm^{-1}); and EC_{OF} and EC_{GW} are the electrical conductivity values in overland flow and groundwater (µS cm^{-1}). EC_{OF} was measured in samples of overland flow collected at the soil surface on hillslopes draining to the stream during the rainfall event (Ribolzi et al., 2018). Because groundwater is supplied to the stream during interstorm flow periods, EC_{GW} was approximated from stream measurements at the beginning of the flood event.

Text S3. *E. coli* concentration monitoring and laboratory analysis

The *E. coli* concentration was monitored at the gauging and sampling station by collecting stream water samples (500 mL) in clean plastic bottles during both base flow and stormflow events with an average frequency of 15 d. However, this sampling frequency was not consistent over the 8 years, which led to a discontinuous time-series of *E. coli* concentration. The water samples were kept in a cool box, and their analysis was carried out within 24 h of collection.

To measure the *E. coli* concentration, we used the standardized microplate method (ISO 9308–3). Each sample was incubated at four dilution rates (1:2, 1:20, 1:200, and 1:2000) in a 96-well microplate (MUG/EC, Biokar Diagnostics) for 48 h at 44 °C. Then, the Ringers' Lactate solution was used for dilution, and one plate was used for each sample. We then noted the number of positive wells for each microplate. The Poisson distribution was used to calculate the most probable number (MPN) per 100 mL. This microplate method has been successfully applied in other studies in the northern Lao PDR (Ribolzi et al., 2016; Kim et al., 2017).

Similar to grab sampling, we collected samples of stream water at the monitoring station using clean plastic bottles and an automatic sampler (Automatic Pumping Type Sediment Sampler, ICRISAT) for the measurement of *E. coli* concentration during 11 flood events. The automatic sampler was triggered by the water level recorder to collect water after every 2 cm of water-level change during the rising of the flood and after every 5 cm of water level change during recession.

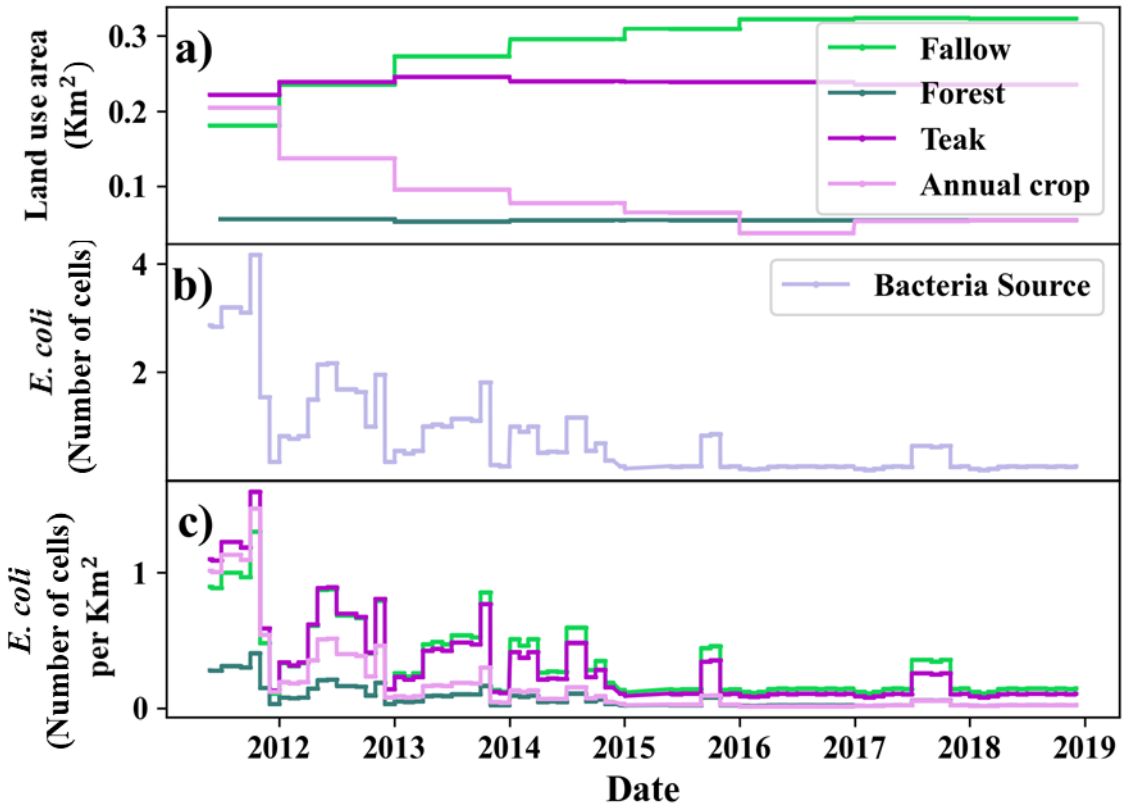
Text S4. Sensitivity analysis and optimization

During calibration processes, it is difficult to optimize a large number of parameters, so we conducted a sensitivity analysis to determine which parameters affect the model output the most. We used Python's open-source library, "*SALib*" (Herman and Usher, 2017), to implement the method of Morris, namely, one-at-a-time (OAT) (Morris, 1991). We used 13 PERLND-associated parameters (**Table S1**) for each land use for sensitivity analysis. As our catchment included four land uses, the total number of parameters was 52. *SALib* performed a sensitivity analysis by varying one variable at one time while keeping all other variables constant. This process was repeated for all variables, and the model output was recorded for each run. The model response in our case was the MSE value between the simulated and predicted surface and subsurface flow. This method of Morris, which is called the OAT method, has been used in many hydrological studies for sensitivity analysis (van Griensvan et al., 2006; Baek et al., 2017; Cho et al., 2012). These parameters were then ranked according to their sensitivity.

After sensitivity analysis, we calibrated the most sensitive parameters using the truncated Newton algorithm (Nash, 1984) provided by the Scipy library (Jones et al., 2001) of the Python programming language. During calibration, we optimized the model parameters. The optimization we chose uses gradient information and optimized the parameters between specific bounds. The bounds for all parameters are given in **Table S2** and were taken from the literature (USEPA, 2000). The optimized values obtained after calibration are given in **Table 3** in the manuscript.

We also conducted optimization based on different objective functions. We used MSE and NSE calculated for simulated surface flow as well as for subsurface flow as an

objective function. During these optimization scenarios, the parameters of the HSPF, which control the surface and subsurface flow, were optimized.



113

114 **Figure S1:** Land-use change and bacteria source from 2011 to 2018 in the Houay Pano

115 catchment, northern Lao PDR: (a) Land-use change, (b) Monthly variation of bacteria

116 source, and (c) *E. coli* source for each land use.

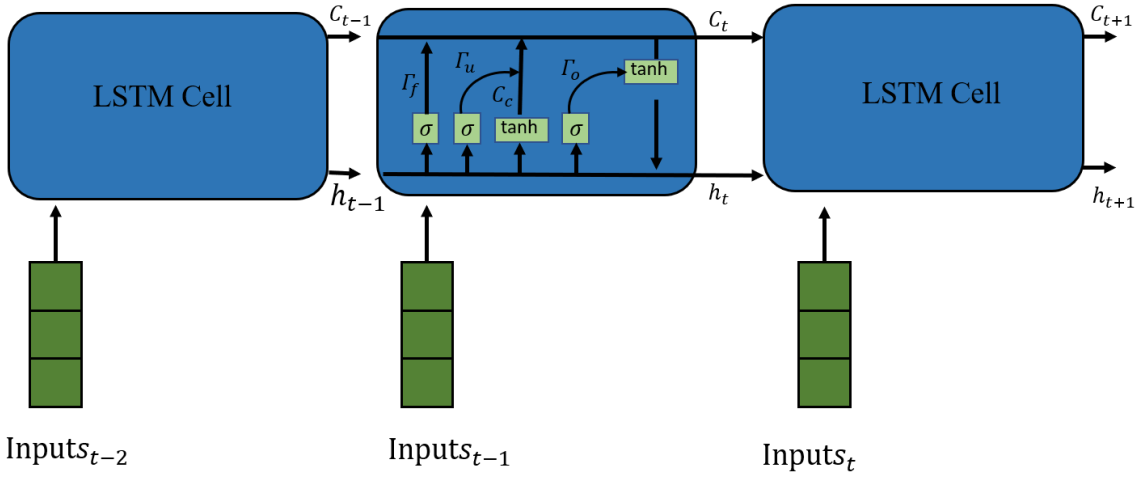
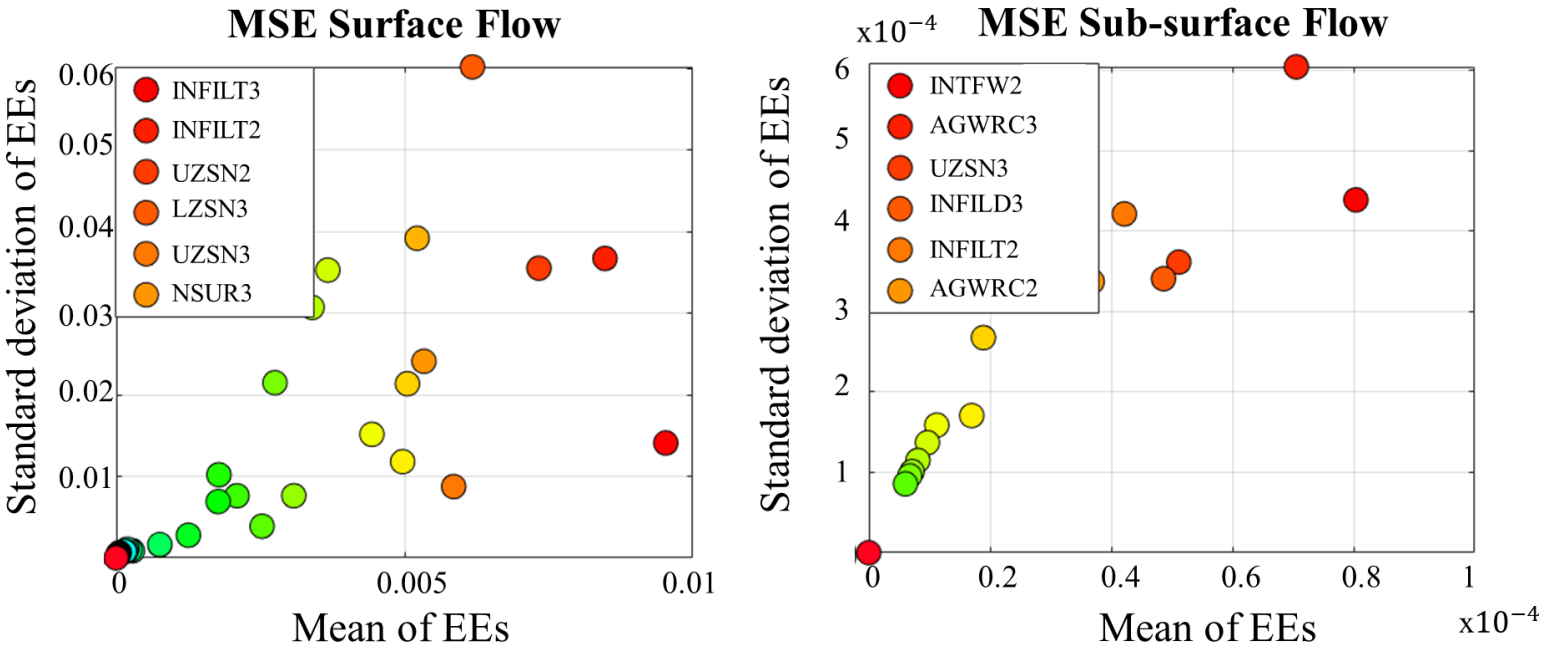


Figure S2: Description of an LSTM layer. An LSTM layer consists of LSTM cells which process information at one time step and generates cell state (c_t) and hidden state (h_t) which are fed to the next LSTM cell. The hidden state is considered as output. The LSTM cell consists of "forget" gate, "update" gate, and "output" gate. σ and tanh represent sigmoid and hyperbolic tangent nonlinearities



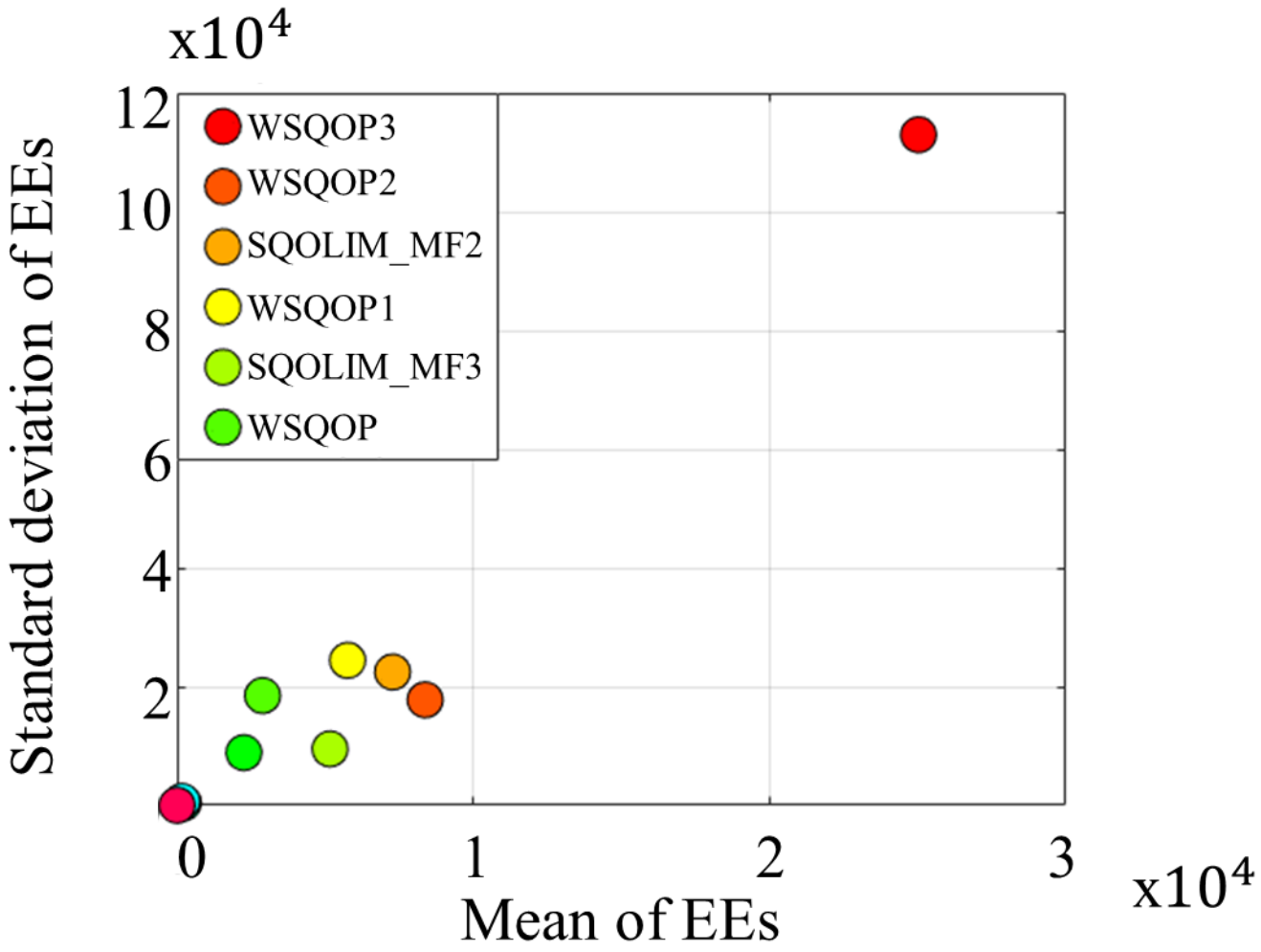
127

128 **Figure S3:** LH-OAT sensitivity analysis of hydrology parameters in HSPF. EEs

129 represent elementary effects. Details of abbreviations are given in Table S1. Boxes in

130 each plot show the five most sensitive parameters. Numbers in legends represent land

131 use; 1: Forest, 2: Teak, 3: Fallow, and 4: Annual crop.



132

133

134 **Figure S4:** LH-OAT sensitivity analysis of *E. coli* parameters in HSPF. EEs represent

135 elementary effects. Details of abbreviations are given in Table S1. Boxes in each plot

136 show the five most sensitive parameters. Numbers in legends represent land use; 1:

137 Forest, 2: Teak, 3: Fallow, and 4: Annual crop.

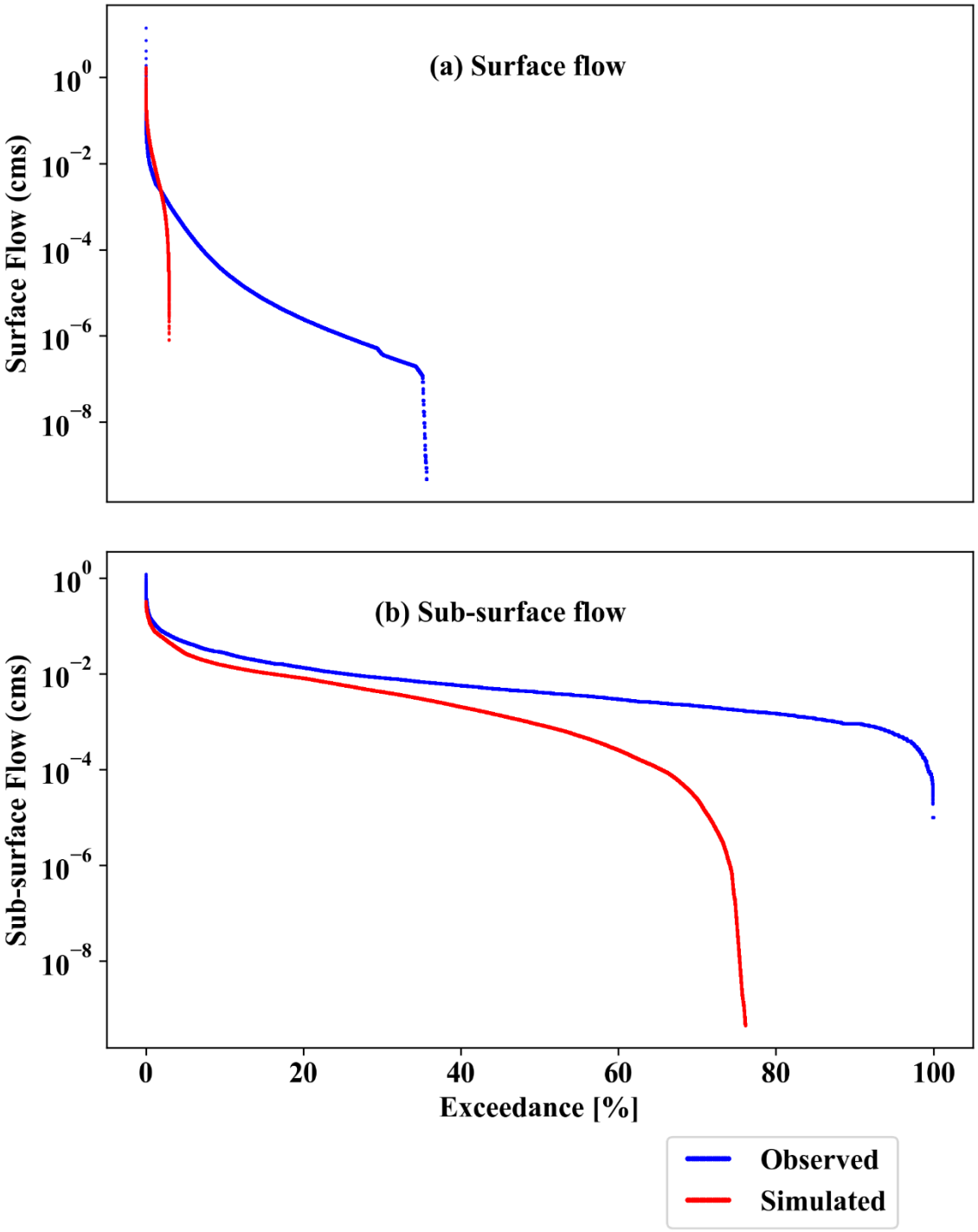


Figure S5: Flow duration curve for surface flow and subsurface flow from HSPF.

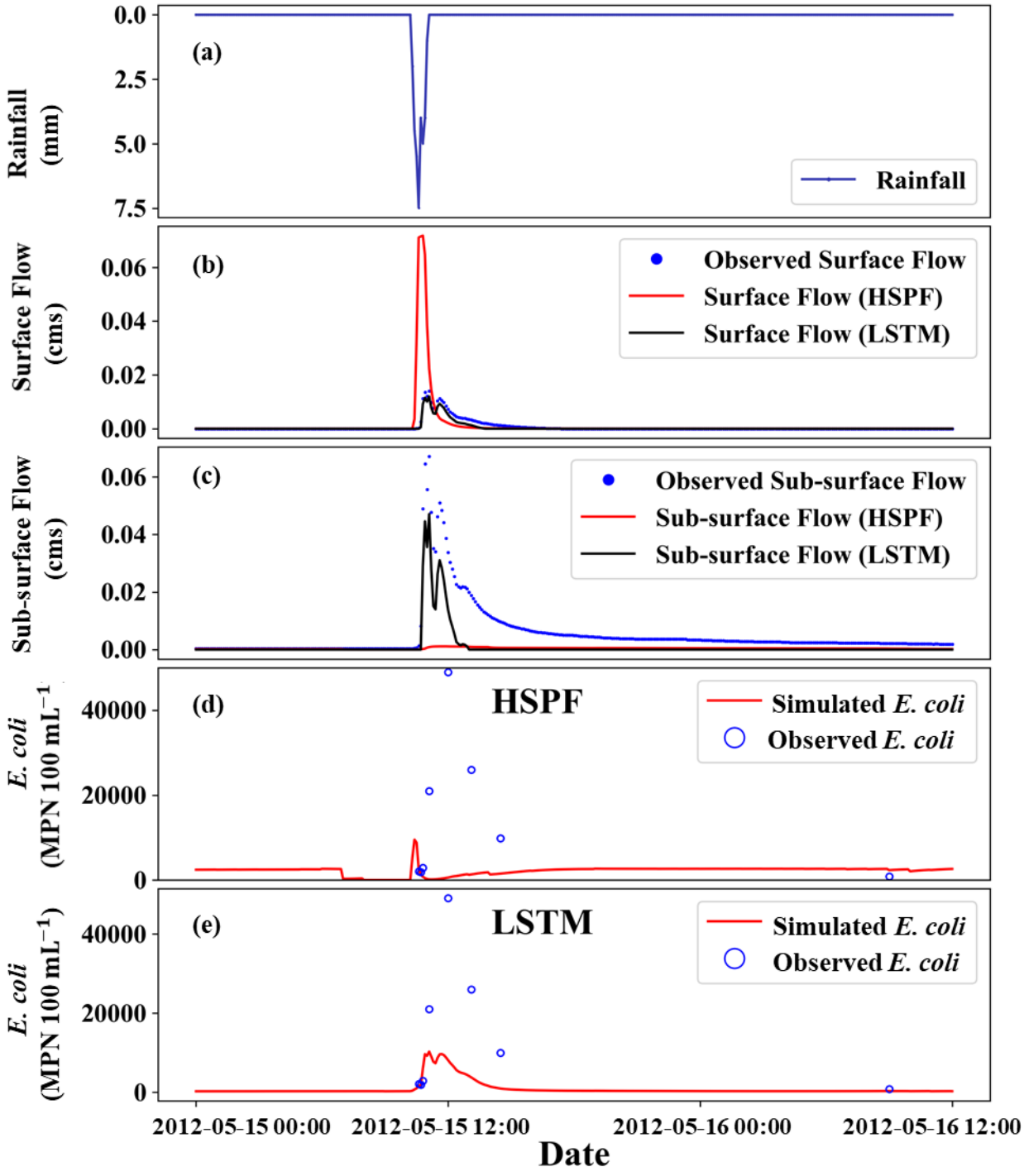
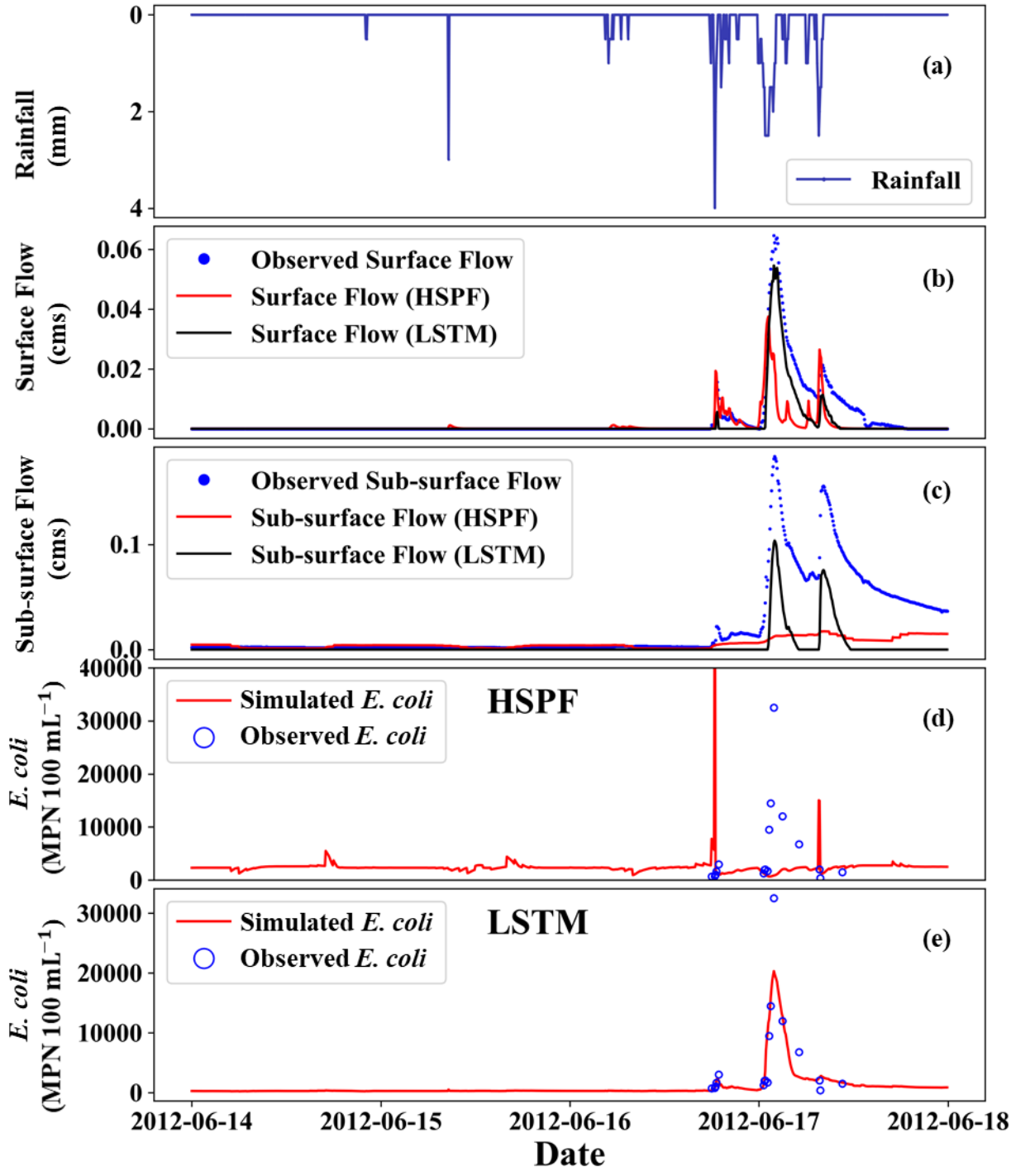


Figure S6: *E. coli* concentration of HSPF and LSTM on May 15, 2012. (a) Observed rainfall, (b) Simulated and observed surface flow, (d) Simulated *E. coli* from HSPF, and (d) Simulated *E. coli* from LSTM.

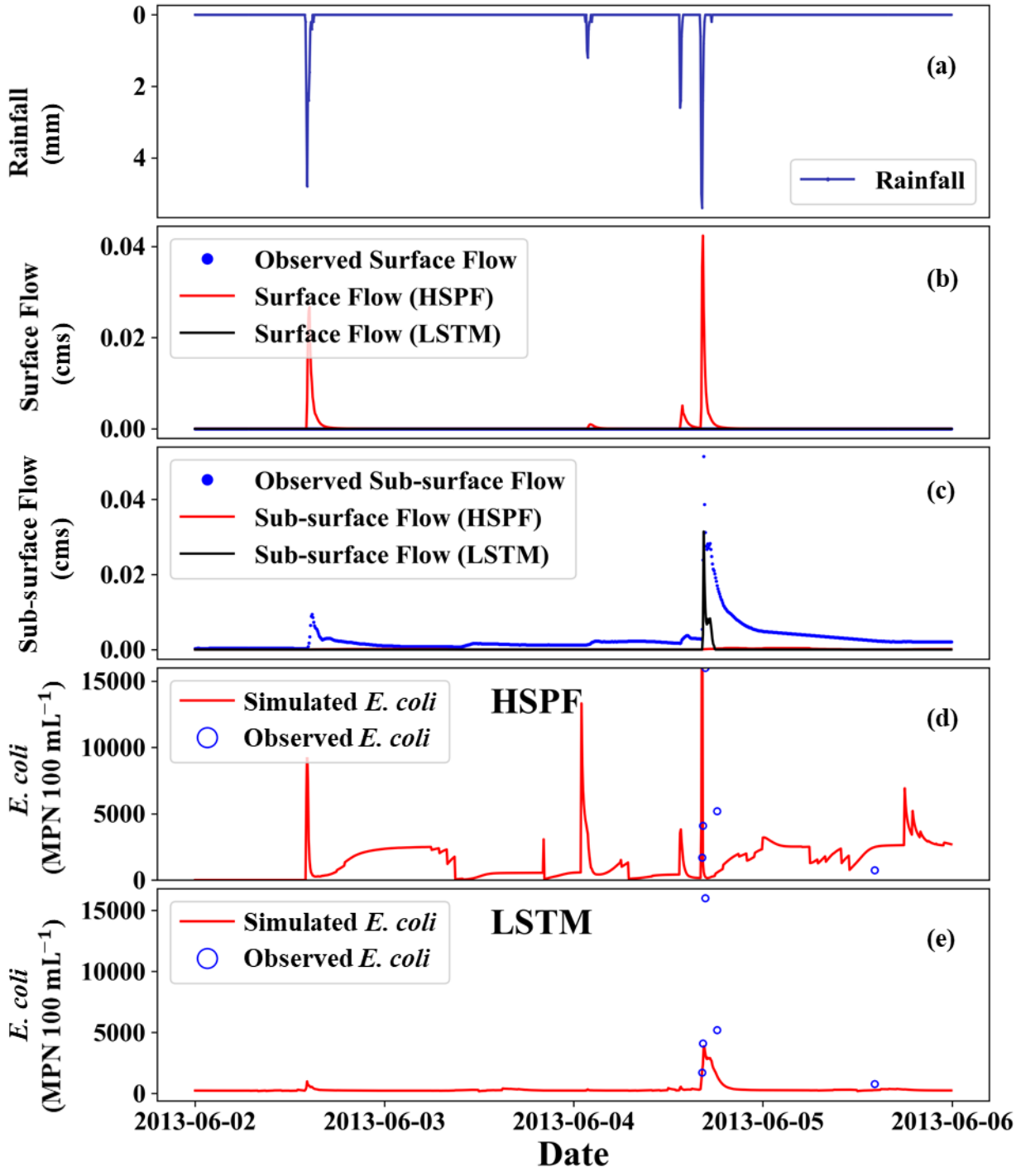


145

146 **Figure S7:** *E. coli* concentration of HSPF and LSTM on June 14, 2012. (a) Observed rainfall, (b)

147 Simulated surface flow, (c) Simulated subsurface flow, (d) Simulated *E. coli* from HSPF, and (d)

148 Simulated *E. coli* from LSTM.

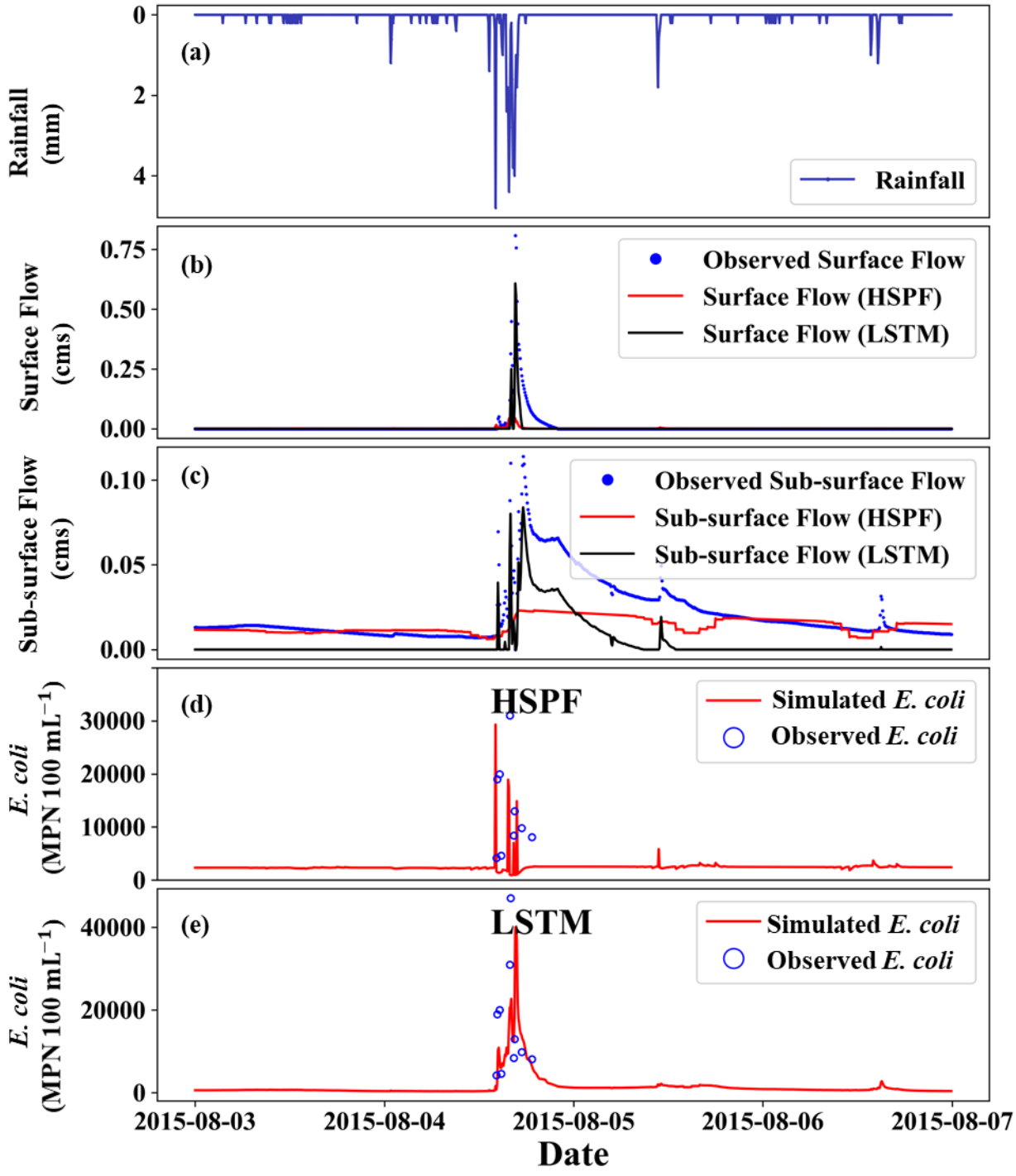


149

150 **Figure S8.** *E. coli* concentration of HSPF and LSTM on June 02, 2013. (a) Observed rainfall, (b)

151 Simulated surface flow, (c) Simulated subsurface flow, (d) Simulated *E. coli* from HSPF, and (d)

152 Simulated *E. coli* from LSTM.



153

154 **Figure S9:** *E. coli* concentration of HSPF and LSTM on August 03, 2015. (a) Observed rainfall,

155 (b) Simulated surface flow, (c) Simulated subsurface flow, (d) Simulated *E. coli* from HSPF, and

156 (d) Simulated *E. coli* from LSTM.

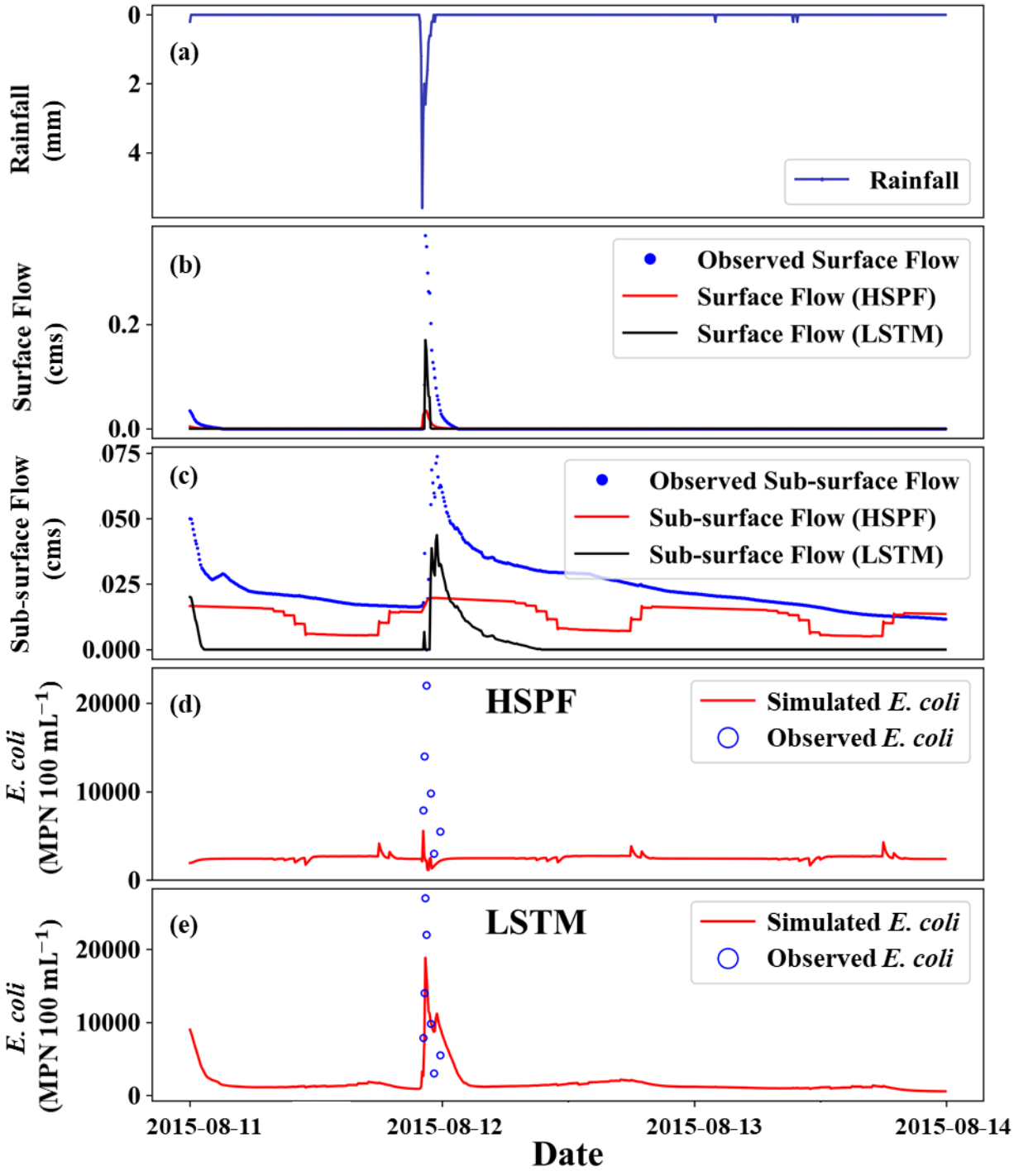
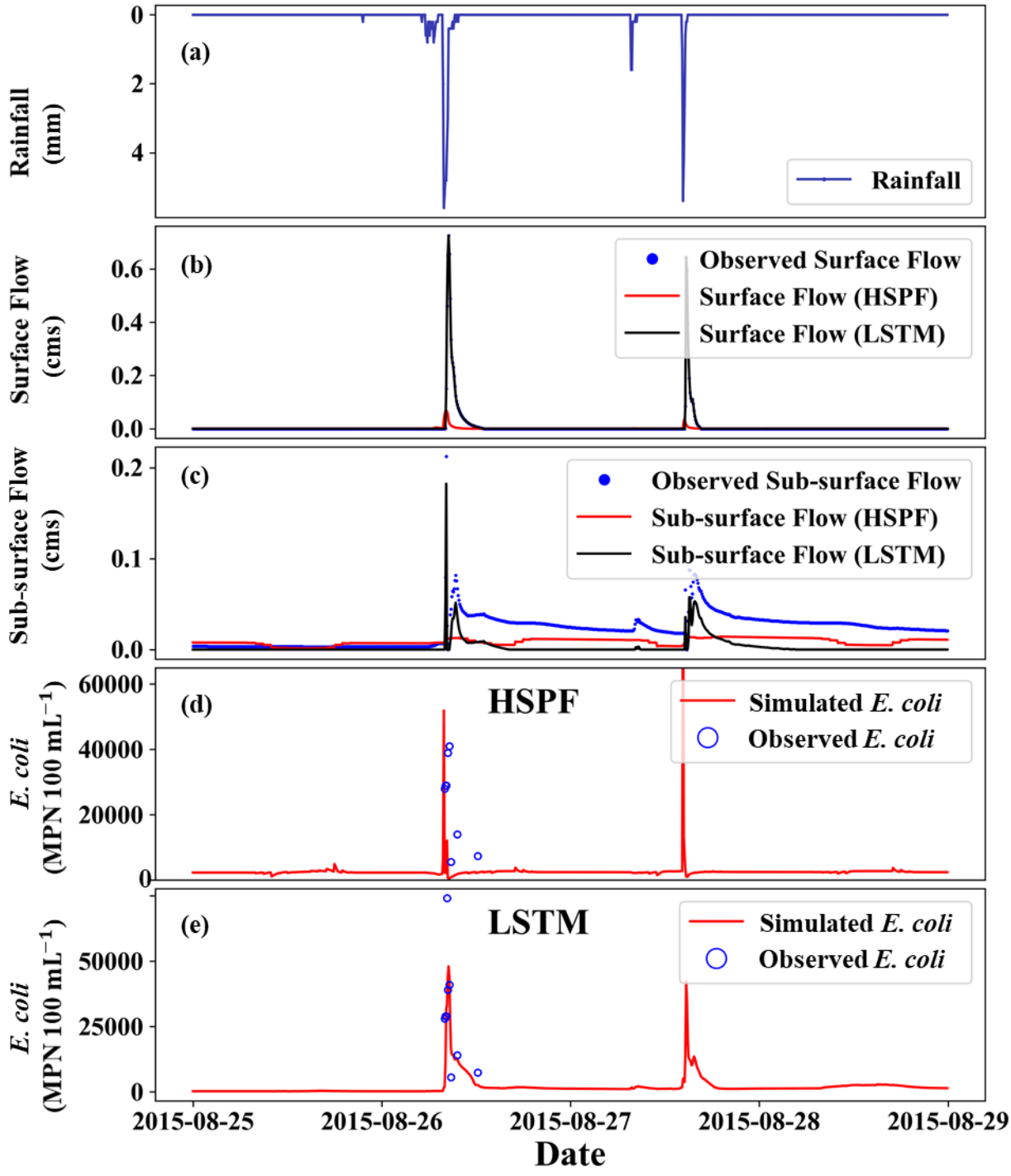


Figure S10: *E. coli* concentration of HSPF and LSTM on August 11, 2015. (a) Observed rainfall, (b) Simulated surface flow, (c) Simulated subsurface flow, (d) Simulated *E. coli* from HSPF, and (d) Simulated *E. coli* from LSTM.



161

162 **Figure S11:** *E. coli* concentration of HSPF and LSTM on August 25, 2015. (a) Observed

163 rainfall, (b) Simulated surface flow, (c) Simulated subsurface flow, (d) Simulated *E. coli* from

164 HSPF, and (d) Simulated *E. coli* from LSTM.

Table S1 Abbreviations of HSPF parameters

Abbreviation	Detailed Name
INFILT	Index to mean soil infiltration rate (inches/hour)
UZSN	Upper zone soil moisture storage (inches)
LZSN	Lower zone soil moisture storage (inches)
NSUR	Manning's n for overland flow plane
INFTFW	Interflow inflow parameter
INFILD	Ratio of max/mean infiltration capacities
BASETP	Fraction of remaining evapotranspiration from baseflow
DEEPFR	Fraction of groundwater inflow to deep recharge
AGWETP	Fraction of remaining evapotranspiration from active groundwater
AGWRC	Base groundwater recession
FSTDEC	first-order decay rate for *E. coli*
THFST	Temperature Correction Coefficient for first-order decay of *E. coli*
SQOLIM	The maximum storage *E. coli* in the surface flow
WSQOP	the rate of surface flow that will remove 90 percent of stored *E. coli* in surface flow per hour.
ACQOP	the rate of accumulation of *E. coli* in surface flow.
AOQC	Concentration of *E. coli* in active groundwater flow
IOQC	Concentration of *E. coli* in Interflow

Table S2 Sensitivity ranking of HSPF parameters for surface and subsurface flow with respect to Mean Square Error. Numbers represent land-use; 1: Forest, 2: Teak, 3: Fallow, and 4: Annual crop

Rank	Surface Flow	Subsurface flow
1	INFILT3	INTFW2
2	INFILT2	AGWRC3
3	UZSN2	UZSN3
4	LZSN3	INFILD3
5	UZSN3	INFILT2
6	NSUR3	AGWRC2
7	LZSN2	UZSN2
8	INFILT4	INFILT3
9	INTFW3	INTFW3
10	INTFW2	INFILT4
11	LZSN4	NSUR3
12	INFILD4	UZSN4
13	NSUR2	INFILD2
14	NSUR4	INTFW1

15	UZSN4	DEEPFR2	172
16	INFILD3	LZSN1	
17	INFILD2	LZSN2	
18	INTFW4	LZSN3	
19	INFILT1	LZSN4	
20	INTFW1	INFILT1	
21	UZSN1	AGWRC1	
22	NSUR1	AGWRC4	
23	LZSN1	INFILD4	
24	BASETP3	DEEPFR1	
25	BASETP2	DEEPFR3	

Table S3 Sensitivity ranking of HSPF parameters for *E. coli* simulation with respect to Mean Square Error. Number in parameter represents land-use; 1: Forest, 2: Teak, 3: Fallow, and 4: Annual crop

Rank	**Parameter**
1	WSQOP3
2	WSQOP2
3	SQOLIM_MF2
4	WSQOP1
5	SQOLIM_MF3
6	WSQOP4
7	SQOLIM_MF1
8	FSTDEC
9	THFST
10	SQOLIM_MF4
11	AOQC4
12	AOQC2
13	AOQC3
14	AOQC1
15	IOQC3
16	IOQC2
17	IOQC4
18	IOQC1

177

178 **References**

179

180 Collins, R., and Neal, C.: The hydrochemical impacts of terraced agriculture, Nepal. *Science of*
181 *the total environment*, *212*(2-3), 233-243, 1998.

182

183 Hochreiter, S., and Schmidhuber, J.: Long short-term memory. *Neural computation*, *9*(8), 1735-
184 1780, 1997.

185

186 Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., and Veith, T. L.:
187 Model evaluation guidelines for systematic quantification of accuracy in watershed simulations.
188 *Transactions of the ASABE*, *50*(3), 885-900, 2007.

189

190 QGIS Development Team.: QGIS geographic information system. *Open source geospatial*
191 *foundation project*, 2016.

192

193 Ribolzi, O., Evrard, O., Huon, S., De Rouw, A., Silvera, N., Latsachack, K. O., ... and
194 Sengtaheuanghoung, O.: From shifting cultivation to teak plantation: effect on overland flow and
195 sediment yield in a montane tropical catchment. *Scientific Reports*, *7*(1), 1-12.
196 https://doi.org/10.1038/s41598-017-04385-2, 2017.

197

198 Ribolzi, O., Lacombe, G., Pierret, A., Robain, H., Sounyafong, P., De Rouw, A., ... and
199 Latxachak, K. O.: Interacting land use and soil surface dynamics control groundwater outflow in

a montane catchment of the lower Mekong basin. *Agriculture, Ecosystems & Environment*, *268*, 90-102. https://doi.org/10.1016/j.agee.2018.09.005, 2018.

Waseem, M., Mani, N., Andiego, G., and Usman, M.: A review of criteria of fit for hydrological models. *International Research Journal of Engineering and Technology (IRJET)*, *4*(11), 1765-1772, 2017.