Dear Reviewers,

Thank you very much for your time involved in reviewing the manuscript and your very encouraging comments on the merits.

Comments:

“The authors proposed an LSTM-based emulator to simulate the ponding process in the drainage system, which is critical to urban flooding study. The emulator is composed of two LSTM models to sequentially simulate node lateral flows and the ponding volume, followed by a correction model. The proposed emulator was successfully applied to a case study and showed superior performances over some simplified versions (e.g., a lumped model using LSMT/CNN). I appreciate the hard work that has been put in by the authors. However, I have the following concerns which might require further revision before the manuscript can be accepted.”

We also appreciate your clear and detailed feedback and hope that the explanation has fully addressed all of your concerns. In the remainder of this letter, we discuss each of your comments individually along with our corresponding responses.

To facilitate this discussion, we first retype your comments in italic font and then present our responses to the comments.

Comment 1:

First of all, I had a hard time following the manuscript. Readability is critical to a renowned journal such as HESS. The current status of the manuscript does not meet the requirement. For example, there are a lot of run-on sentences. A rule of thumb is that the length of a sentence does not exceed two lines. Coherence is also an issue. Many sentences are 'loosely' connected in a logical sense. It would be a pity if the message is not clearly communicated while so much work has been done. I suggest the authors further greatly revise the language (a professional English editor might help in this case).

Response 1:

Thanks for your suggestion on improving the accessibility of our manuscript. We have revised the language by a professional English speaker. We have split the run-off sentences, and revised the logically incoherent sentences.

Comment 2:

The second issue is associated with the model CR of the LSTM-based emulator (btw, what is CR abbreviated for?). I don't quite understand the descriptions of the model CR (i.e., L153-166). Neither Figure 6 is illustrative to me. Do the monitoring data refer to the measured lateral flows at the monitored nodes? Is the correction model trained on pairs of simulations and monitored measurements or based on a pre-trained mapping (i.e., using transfer learning)? Please specify.

Response 2:

Thanks for your comment. CR is abbreviated for correction of the runoff process. Monitoring data refer to the measured water depths at the monitored nodes and flows at the monitored pipelines.

Model CR is designed to update the runoff process in the primary LSTM-based model. The correction has two steps: training and updating. Firstly, the model CR is trained based on a pre-trained mapping from X to Y (as shown in Fig.6). Then, it is updated on pairs of measured rain data, monitored water depths and flows.

We have adjusted Figure 6 in our paper. The related contents are also provided below for your quick reference. See L135, L146-147 and L150-152 for details.

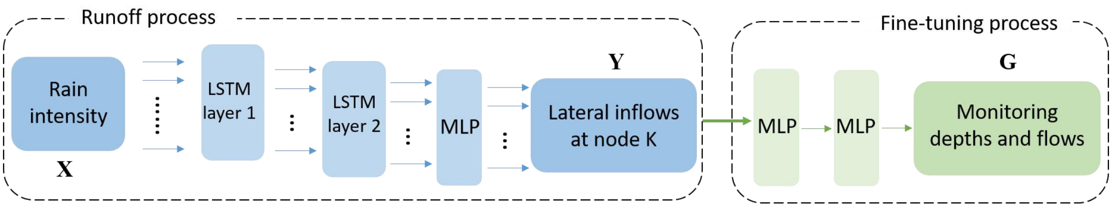

Figure 6:The architecture of Model CR. (MLP -- Multi-Layer Perception)

Comment 3:

The last concern is the mass balance of the emulator. Though not an expert in urban drainage systems, I consider that the mass conversation plays a key role in balancing the water exchanges between nodes. Does the proposed LSTM model account for that? If not, please specify the reason for not doing this.

Response 3:

Thanks for your comment. The proposed LSTM model does account for the mass conversation between nodes. The model is trained using data simulated by mobilizing the SWMM model. The hydrodynamic model mainly simulates the flow of runoff and external flows in the pipes, channels and etc, including overflow, outflow and transmission of the pipe network system. The water exchanges between nodes have been covered in the simulation process.

Comment 4:

Other minor edits:

L37-40: The authors point out the importance of the dataset. I'm wondering whether the author performed a sort of convergence test to evaluate how much data is sufficient for the proposed LSTM emulator training.

Response 4:

Thanks for your comment. The related supplement has been added in our paper and are also provided below for your quick reference. See L210-216 for details.

“In general, a small training set normally leads to poor approximation effect. Thus a convergence test was performed to evaluate how much data was required for the proposed LSTM-based model to obtain the desired approximation effect. The model performances using different sizes of training data were compared, as shown in Fig. 9. When the data size was reduced to 2/3 of the origin volume, the model performance fell down to 90% of the original. And if the data size was halved, less than 80% of the origin model performance remained. ”

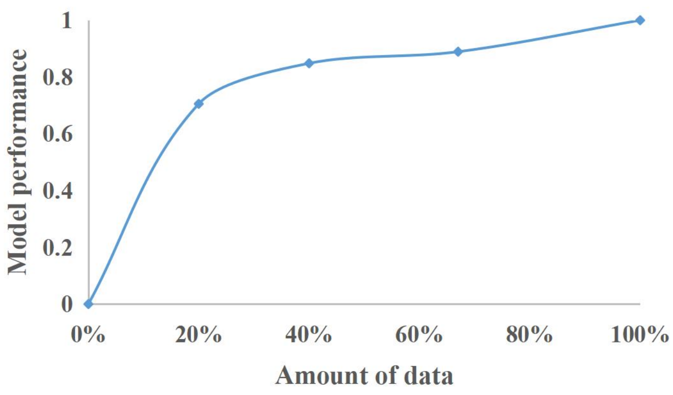

Figure 9: The learning curve which describes the relationship between model performance and data volume.

Comment 5:

L50: 'not discussed' --> 'not explored'; 'not available' --> 'not feasible'

L77: 'influencing' --> 'influential'

L195: 'tb and ta is' --> 'ta and tb are'

L226: I like the usage of hyperopt here.

L338: 'In a summary' --> 'In summary'

Response 5:

Thanks for your suggestion on improving the accessibility of our manuscript. These questions have been revised one by one.

Comment 6:

L82: 'MAE, MSE, CC, NSE' --> We usually put full names before abbreviations

Response 6:

Thank you for your comment. The related content have been added in our paper. See L85-87 for details. The related contents are also provided below for your quick reference.

“(MAE -- Mean Absolute Error, MSE -- Mean Squared Error, CC -- Correlation coefficient, NSE --

Nash-Sutcliffe efficiency coefficient)”

Comment 7:

Figure 2 caption: '... test process in the runoff process' --> '... test procedures in developing the LSTM-based runoff emulator'. Also, many captions are too brief to provide enough information about these complicated figures.

Response 7:

Thank you for your comment. The related content have been revised in our paper. See L96, L114, L135, L206, L306-307, L354-356, L363-364 for details. They are also provided below for your quick reference.

Figure 2: Figure 2: The training, validation, and testing procedures used when developing the LSTM-based runoff emulator. (MAE -- Mean Absolute Error, MSE -- Mean Squared Error, CC -- Correlation coefficient, NSE -- Nash-Sutcliffe efficiency coefficient)

Other captions have also been modified as follows:

Figure 3: The network structure of the flow confluence process (for a single node).

Figure 4: The error transmission during training from the runoff process to the flow confluence process.

Figure 5: Model correction system. CR is abbreviated for correction of the runoff process.

Figure 8: A demonstrative example to show the effect of adding white noise.

Figure 12: Comparison between the predicted ponding volume and simulation from the hydrodynamic model at the selected nodes in several testing rainfall events randomly chosen.

Figure 16: Schematic diagrams of different network structures for comparison. (a) The same runoff process in model A and B. (b) The multi-target learning in the flow confluence process of model A marked light blue. (c) The flow confluence process in model B marked dark blue. (d) The LSTM in model C. (e) The CNN in model D.

Figure 17: Comparison of model performance on the ponding volume forecasting. The results of the proposed model A are compared to those obtained from models B, C, and D.

Comment 8:

Figure 3: For each of the two emulated processes, is only one LSTM used for all nodes? Or, is a separate LSTM used for each node?

Response 8:

Thanks for your comment. For each of the two emulated processes, a separate LSTM is used for each node.

Comment 9:

L91-93: That's a super long sentence and there are a lot!

Response 9:

Thank you for your comment. The long sentence has been revised and is also provided below for your quick reference. See L89-91 for details.

"The flow confluence process is set up in the same manner as the simulation process of a hydrodynamic model (e.g., the SWMM model). If we compare the urban drainage system to a black box, only the lateral inflows at each node and outflows from the outlets enter and leave the system, respectively (Archetti et al., 2011)."

Comment 10:

L100-102: Are the classification module and OUT_MODULE also two MLPs?

Response 10:

Thanks for your comment. The 'CLASSIFICATION MODULE' and 'OUT MODULE' are two separate MLPs as the outputs for two tasks respectively.

Comment 11:

L105-106: I don't understand which layer in the LSTM module is shared by the classification and out modules.

Response 11:

Thanks for your comment. We have revised our paper. The related contents are also provided below for you quick reference. See L105-108 for details.

"Moreover, the multi-task learning has a hard parameter sharing mechanism, which effectively alleviate the over-fitting of the model. The parameters in the 'LSTM_MODULE' (including the parameters of the LSTM layers, batch normalization layers, activation functions, etc.) are shared by the 'CLASSIFICATION_MODULE' and ′OUT_MODULE'."

Comment 12:

L116-119: To evaluate the impact of the gaussian filter, is there a comparison between the current emulator and one without the gaussian noising procedure?

Response 12:

Thanks for your comment. We have revised the related contents and provided them below for your quick reference. See L97-100 for details.

“As illustrated in the pink block in Figure 3, a Gaussian layer is added after the input layer in the flow confluence process during training. The gaussian layer serves as a filter to compensate for the inaccuracy of the prediction (by the hydrodynamic model) in the runoff process. The model is trained to minimize the differences between the predictions (from the neural network, i.e. the output from the runoff process) and the simulations (from the hydrodynamic model).”

Comment 13:

Eqs(1)-(4): I suggest moving the calculation of the error term to the appendix to improve the readability.

Response 13:

Thank you for your comment. However, we believe that this part is very important in our model. The error is to explain the performance of the runoff process, and to compensate for the difference between the input of the flow confluence process in the actual application and the simulation data used in the training. So we think that it is better to put the calculation of the error in the text.

Comment 14:

Eq.(5): What is 'lg'? Please use 'log' if you mean logarithm operation.

Response 14:

Thank you for your comment. We have revised Eq.(5) in our paper. See L182 for details. It is also provided below for your quick reference.

$$q = \frac{167A(1 + C\log P)}{(t + b)^n} = \frac{1600(1 + 0.846\log P)}{(t + 7.0)^{0.656}} \quad (5)$$

Comment 15:

L197: Is Pilgrim & Cordery a reference? If yes, please provide the year.

Response 15:

Thanks for your comment. The related reference has been added to our paper. The related contents are also provided below for your quick reference. See L190-191 for details.

“The Pilgrim & Cordery was a method to count the historical rainfall data, and deduce the rainstorm pattern from it (Pilgrim and Cordery, 1975).”

Comment 16:

L229-231: missing subjects of the two sentences.

Response 16:

Thank you for your comment. We have revised these two sentences in our paper. See L237-238 for details. They are also provided below for your quick reference.

"Then, it received a random combination of hyper-parameters from the search space to do the training and returned the training loss. It used the built-in search algorithms like the tree of Parzen estimators (TPE) to determine the following hyper-parameters combination."

Comment 17:

Table 3: What are the optimal hyperparameters of the MLP used for model CR? i.e., the number of neurons in each layer and the number of hidden layers. How about the hyperparameters of the classification and out modules?

Response 17:

Thanks for your comment. We have revised Table 3 in our paper. The related contents are also provided below for your quick reference. See L234-236 for details.

Table 3: Hyper-parameters configuration in model setup and correction processes.

	Hyper-parameters	Model CR		Flow confluence process
		Runoff process	Fine-tuning process	
Model setup	Normalization	Z-score	Z-score	Min-Max
	Batch size	150	150	150
	Epoch	300	300	300
	Learning rate	1e-2	5e-3	1e-2
	Optimizer	Adam	SGD	SGD
	LSTM hidden layer neurons	16	-	256
	MLP hidden layer neurons	16	1536/3072	256/128*
	LSTM layers	2	-	4
	MLP layers	1	2	2*
Model correction	Learning rate	1e-4		5e-5
	Optimizer	Adam		SGD

Note: * means to set the hyperparameters of 'CLASSIFICATION_MODULE' and 'OUT_MODULE' in the flow confluence process to the same values.

"The column 'Fine-tuning process' in Table 3 lists the optimal hyperparameters of the MLP used for model CR. The number of MLP hidden layers is 2, and the numbers of neurons in each layer are 1536 and 3072 respectively. Besides, the hyperparameters of the classification and out modules in the flow confluence process are set to the same values. The number of MLP hidden layers is 2 and the numbers of neurons in each layer are 256 and 128 respectively."

Comment 18:

L274: Why are these six nodes selected? (also shown in Figure 9)

Response 18:

Thanks for your comment. The related supplement has been added to the paper. The contents are also provided below for your quick reference. See L276-280 for details.

"The six nodes were selected according to their severity of ponding and the spatial distribution to show the difference in model performance between nodes. Three of them were chosen because their percentages of samples where ponding occurred in the training set were less than 50%, and others were opposite. For example, the percentage of Node 238 was 18.33%, while that of Node 95 was 98.6%. Moreover, Fig. 10 marks their locations."

Comment 19:

Figure 10: It is the emulated ponding volume before the model correction or CR, right? If yes, why is it different from the lines labeled by 'Before updating' in Figure 11?

Response 19:

Thanks for your comment. The ponding at the selected nodes in Figure 12 occurred in some designed rain events randomly selected from the test set. However, the lines labeled 'Before updating' in Fig.15 are drawn in the measured rainfall event No.5.

Comment 20:

L336-341: Should these sentences be grouped into one paragraph?

Response 20:

Thank you for your comment. We have grouped these sentences into one paragraph. See L341-345 for details. They are also provided below for your quick reference.

"Fig. 15 plotted the specific ponding process at the selected nodes in the measured rainfall event No.5 predicted by the corrected model compared to those anticipated before correction. Fig. 15 showed that the revised model performed better at all the selected nodes in terms of not only starting and ending time but also the process of ponding than the uncorrected model. The prediction by the corrected model had a high consistency with the measured ponding volume at each node in the figure, which proved the reliability of the approach to introducing the measured monitoring data to correct this LSTM-based model as well."

Comment 21:

Figure 15: combining (a) and (b)?

Response 21:

Thank you for your comment. We have revised Figure 17 in our paper. See L362-364 for details. It is also provided below for your quick reference.

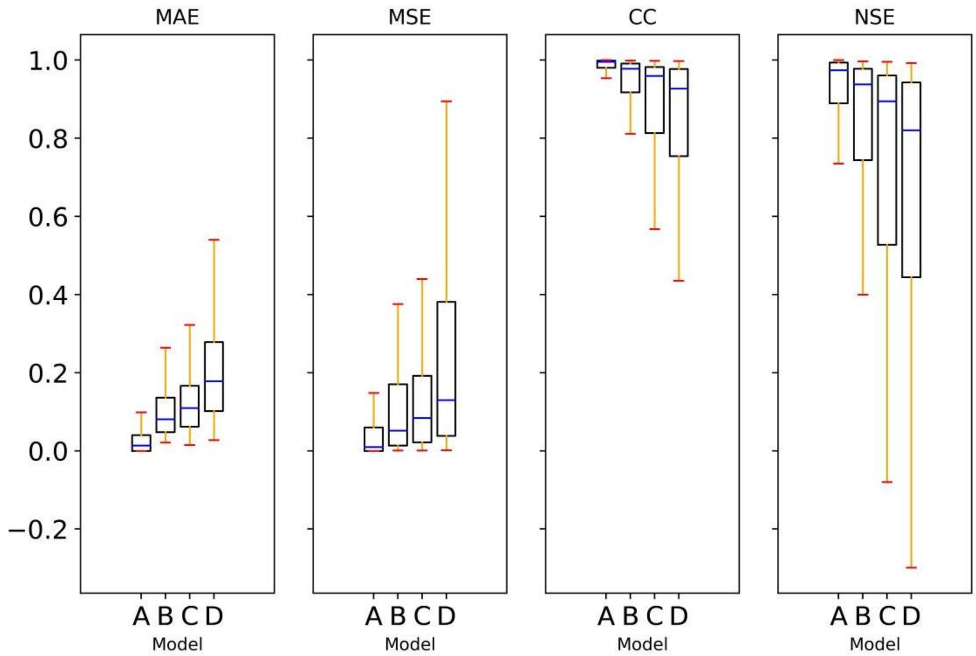

Figure 17: Comparison of model performance on the ponding volume forecasting. The results of the proposed model A are compared to those obtained from models B, C, and D.

We would like to take this opportunity to thank you for all your time involved and for this great opportunity for us to improve the manuscript. We hope you will find this revised version satisfactory.

Sincerely,

The Authors