Responses to comments by referee #1 (Konstantinos X. Soulis)

We would like to thank the referee, Konstantinos X. Soulis, for the careful and thorough reading of this manuscript and the thoughtful comments and constructive suggestions, which help improve this manuscript's quality. Our response follows:

Reviewer #1:

1) *The hydrological model description isn't sufficient neither clear enough. The model used is described as semi-distributed. As I can understand by reading the manuscript, in reality it is a simple, cascading buckets, lumped model applied in the entire watershed as a whole with the same lumped parameters values for the entire watershed. Probably the model is described as semi-distributed because it involves different representations for the various parts and processes (rural, urban, LID) but this isn't accurate. (please see the relevant general comment) (page 6, line 189)*

Response: We appreciate this feedback. An additional Section 2.3 SUPERFLEX model framework was created for a more detailed introduction about SUPERFLEX and its application. Besides, some supplementary descriptions about the model used in this research were added in Section 3, Methodology, in which the model was described as a semi-distributed model with two hydrological response units (HRUs) for urban and rural surfaces, respectively. Therefore, there are two different hydrological systems for urban and rural areas separately. That's why it's called as a semi-distributed model.

This model is similar to the FLEX-Topo model (which was introduced in the added Section 2.3 SUPERFLEX model framework). The FLEX-topo model is also a semi-distributed model with three HRUs (plateau, hillslope, and wetland). Inspired by FLEX-topo, this model was created by separating the hydrological systems in urban and rural surfaces of one catchment.

2) *A very short description of the main characteristics of this model would be helpful. (page 6, line 190)*

Response: We thank you for this suggestion. As mentioned before, an additional Section 2.3 SUPERFLEX model framework was created for a more detailed introduction about SUPERFLEX and its application.

3) *was? (page 7, line 205)*

Response: We thank the reviewer for pointing this out. It was revised.

4) *On what basis? How? (page 7, line 214)*

Response: This is about using expert knowledge to increase the efficiency of parameter calibration. The sentence in the article was rewritten as "For parameter calibration, the initial range of each parameter was given roughly based on the experience and information about catchment properties according to the physical realism of these parameters (Breuer et al., 2003; Gharari et al., 2014)."

It should be mentioned that most of the parameters in SUPERFLEX model have physical meanings, just like the popular TOPMODEL. For example, D (Precipitation distribution factor for Su) indicates the area ratio of unsaturated zone in rural areas or green (permeable) surfaces in urban areas. Sumax (Maximum unsaturated storage) could be estimated from the depth of soil layer and porosity. Therefore, a pair of relatively broad preliminary limitations were given based on the empirical value for more efficient parameter calibration and to prevent unrealistic calibration results.

Breuer, L., Eckhardt, K., and Frede, H. G.: Plant parameter values for models in temperate climates, *Ecological Modelling*, *169*(2-3), 237-293., 2003

Gharari, S., Hrachowitz, M., Fenicia, F., Gao, H., and Savenije, H. H.: Using expert knowledge to increase realism in environmental system models can dramatically reduce the need for calibration, *Hydrology and Earth System Sciences, 18*(12), 4839-4859, doi:10.5194/hess-18-4839-2014, 2014

5) This isn't clear. I cannot understand what has been actually done. (page 7, line 214)

Response: This is about using Monte Carlo methods for parameter calibration. After the decision of calibration limitations (as mentioned in the last response), the Monte Carlo sampling method was used to generate random pair of parameters between the limitations. In this research, the semi-distributed model was created by testing the model performance from six-bucket models to eight-buckets models. As more complex models have more parameters, the same number of samples could be seen as "unfair". Therefore, more complex models with more parameters were tested with larger numbers of parameter sample sets to ensure the calibration scale as "fair" as possible.

6) Some of the symbols are not described (e.g. Su, Sf, Sh, Ss). Subscripts in the parameters symbols would be helpful (the symbols should be presented in the same way as in the mathematical expressions and parameters and components description (e.g. Italics with subscripts) (page 8, line 223)

Response: It is about some symbols in Table 2. Actually, Table 2 was introduced after Figure 3 Schematic figure of semi-distributed model structure. The bucket names and their symbolic expression (Su, Sf, Sh, Ss) were introduced under Figure 3. The name of Table 2 was revised as Parameters and model components in the semi-distributed *model as shown in Figure 3*, which may help to avoid confusion.

7) Unclear. Please make clear what was done (specific process / steps) in a simple way. At some point at the beginning, a short description of the model type and characteristics should be provided (e.g. semi-distributed, cascading bucket model, event based? etc.) (page 7, line 203)

Response: The first piece of advice is about model setup. An additional paragraph about the strategies used in model setup was added in Section 3.2.1 Semi-distributed model setup. For the second suggestion about the description of model, as mentioned before, an additional section 2.3 SUPERFLEX model framework was created for a more detailed introduction about SUPERFLEX and its application.

8) Also unclear. What do you mean by the term verified NSEs and what by the phrase "calculated as accuracy indexed" (page 7, line 209)

9) How the "variance of the verified NSEs" is a measure of precision? (page 7, line 209)

Response to 8 and 9: These involve the specific procedure of model structure selection. As mentioned in the revised manuscript and the responses above, 1) the Monte Carlo sampling method was used to generate pairs of parameters, and 2) both the Nash-Sutcliffe Efficiency (NSE) and correlation coefficient (R2) were used to evaluate the model performance. Therefore, each pair of data would get two NSE and R2 values as calibration and validation results, respectively.

To obtain a model structure with both stable (high precision) and outstanding (high accuracy) performance, the optimal, mean, and variance of those verified NSE and R2 were examined to evaluate the accuracy and precision of model structures, as the table below.

Model Num.	Accuracy index					Precision index
	The R^2 of calibration	Optimal verified R^2	*The NSE of calibration*	Optimal verified NSE	Mean of verified NSE	Variance of verified NSE
M01A	*0.882*	0.865	*0.723*	0.735	0.117	0.399
M01B	***0.897***	**0.873**	***0.673***	**0.748**	**0.297**	**0.240**
M02	*0.863*	0.871	*0.712*	0.749	0.155	0.424

M03A	*0.902*	0.850	*0.721*	0.706	-0.056	0.758
M03B	*0.815*	0.855	*0.656*	0.730	0.039	0.587
M04	*0.878*	0.883	*0.714*	0.744	0.036	0.606

Symbols: Boldface row indicates the selected model structure; Italic columns indicate the calibration results.

Since the trivial details is quite complex and of little importance to the main content of the article, the specific procedures was neglected in the manuscript to avoid unnecessary confusion of readers.

10) Below, there is another figure (Fig. 4) with additional structures for various LID scenarios. It should be mentioned here that this structure corresponds to the current case without LID and that this isn't the general model. Please try to provide a more comprehensive description of the model. (page 9, line 229)

Response: We appreciate this insightful suggestion. An additional paragraph was added at the start of Section 3.2 Hydrological model, to distinguish the different model setup methods between current condition and assumed scenarios (urbanization and LID).

11) Please explain this. In table 2, D is the precipitation distribution factor. It is unclear what you mean with that. (page 9, line 242)

Response: Thank you for pointing this out. The description of D was clarified as the Precipitation distribution factor for Su in Table 2.

12) Do you mean the new strucrures presented in Figure 4? (page 10, line 250)

Response: Thanks for the feedback. It was revised by adding "of the original semi-distributed model" in the manuscript. At the end of this paragraph, Figure 4 was introduced as the final schematic model figure of four LID modules.

13) Considering that this model structure is empirical, some more information on how the values of all these parameters that do not have a clear physical basis were specifically assumed, on what base and by which source (for each parameter). Transferring empirical parameters values between different models is challenging. (page 10, line 251)

Response: We thank you for this valuable suggestion. As mentioned in response (4) above, most of the parameters in SUPERFLEX model have physical meanings. And it is feasible to estimate those parameters based on relevant literature, realistic field test results, and data from local government files. Estimating parameter based on prior information has been applied in many studies, especially for this type of ungauged basins (assumed LID implementation). Due to space restrict, the specific parameter estimating procedure was deleted in the manuscript. But an example of parameter estimating for bioretention cells are provided here:

There are five parameters in the bioretention model component, the precipitation distribution factor (D_{LID}), the ratio of drainage area to construction area (A_R), the maximum interception depth ($I_{max,B}$), the maximum water storage depth in soil layer (S_{umax}), and time lag coefficient of bioretention cells (T_{lagB}).

Both of the precipitation distribution factor (D_{LID}) and the ratio of construction area to drainage area (A_R) depends on the concrete LID implementation plan, which two could be adjusted to fit different LID scenarios.

The maximum interception depth ($I_{max,B}$) indicated the interception capacity of bioretention cells. Li et al. (2009) found that the intercept depths for six bioretention facilities ranged from 0.6 to 4.6 mm in Maryland, U.S. A relatively good vegetation condition of the bioretention is assumed in this project with the 3.5 mm interception capacity. To adapt this parameter into the urban module, the assumed

interception capacity should be multiplied by the precipitation distribution factor (D_{LID}) and be divided by the ratio of construction area to drainage area (A_R), as the final parameter, $I_{max,B}$;

As for the maximum water storage depth $S_{umax,B}$, according to "SARA LID Guidance Manual", 2 to 5 feet (0.6 to 1.2 m) soil media depth is recommended for the bioretention design. And the average soil media depth of the six typical bioretention facilities in the Maryland is referred as 0.84 m (Li et al. 2009). Considering these two recommended values, the depth of soil media layer is presumed as 0.85 m in this project, and an empirical soil porosity is chosen as 0.35 since the moderately permeable condition of local soil. Therefore, the water storage capacity for the bioretention is supposed as 300 mm. Then, the water storage capacity should multiply with the precipitation distribution factor (D_{LID}) and be divided by the ratio of construction area to drainage area (A_R), which is the estimated value of parameter $S_{umax,B}$;

Finally, according to a field test by Hunt. (2008), the peak flow of the bioretention cell could be delayed by 3 hours. Therefore, T_{lag}, the number of the delayed time interval, was assumed as 13 (3 (hours) * 2 (intervals per hour) * 2 (symmetric equation) + 1 (starting point)), to fit in the mathematical expression of delay.

14) I was expecting a paragraph concerning model calibration and validation methods, data, hydrometric stations used etc. and a paragraph describing the model application, the data used, modeled events characteristics etc. and the analysis of the results. (page 12, line 265)

Response: For model calibration, validation, and data used, an additional paragraph about the strategies used in model setup was added in Section 3.2.1 Semi-distributed model setup. For model application, an additional Section 2.3 SUPERFLEX model framework was created for a more detailed introduction about SUPERFLEX and its application. For the model performance, Figure 5 was updated as individual peak events were compared during both non-flood and flood seasons.

15) Please check this! Is it exactly the same? Really strange as it concerns both the calibration and the validation periods. (page 12, line 272)

Response: We thank you for your careful reading. The typo was revised from 166 to 160 mm.

16) peaks? (page 12, line 276)

Response: The word was revised according to your suggestion.

17) Is this also a calibration parameter? Does it represent the permeable part of the watershed? It seems that a parameter like this could be estimated by the percentage of impermeable areas or something like this with reasonable accuracy in case that this parameter has the physical meaning that I assumed above. (page 12, line 278)

Response: It is about the parameter D (precipitation distribution factor for the unsaturated zone). Your perception is correct. But in this research, all the parameters for current condition (since hydrological data are available), including D were obtained from calibration. Those parameters in urbanization and LID implementation scenarios were estimated (since no measured hydrological data).

18) In figure 5 seems that the difference is mostly on peak flows, base flow seems to be similar in both cases. (page 12, line 286)

Response: The difference in base flows was indeed insignificant in this figure. This sentence was revised.

19) Isn't it relatively high for urban areas? (page 12, line 288)

Response: It is about the parameter D (precipitation distribution factor for the unsaturated zone) in urban areas, which is 0.83 according to the parameter calibration results. This result can be reasonable as there are some green spaces in urban areas, and there are more in large areas of sub-urban regions.

20) *It would be interesting to also have a figure presenting in detail a storm event, similar to figure 7, in order to be able to see the performance of the model at detailed time scales, as the following discussion puts emphasis on specific events. (page 13, line 290)*

Response: A new version of this Figure 5 could be found in the revised manuscript.

21) *I cannot see this in the figure. It seems like peaks timing is the same in all scenarios. Also I cannot differentiate the rural runoff from scenario A and current situation. It seems like the lower peak in scenario A is the result of lower urban peak flow for some reason. (page 16, line 337)*

Response: Section 4.2 Urbanization influences on basin runoff was rewritten, and Figure 6 was updated with the zoom-in figure. For the comments about the rural and urban sub-flows, it was true that urbanization scenario A and current situation share the same rural sub-flow. It was because the larger areas of urban grey (impermeable) surfaces and less urban green surfaces in urbanization scenario A. This change of urbanization scenario A moved part of the runoff of the latter summit II forward and superimposed on the faster summit I, which cause the decrease of the latter summit II (and increase of summit I). Therefore, the maximum peak decrease, which contributed by the latter summit II.

22) *This is controversial. It could be the case by coincidence in very specific situations. (page 16, line 357)*

Response: It is about the urbanization influences in flood season. Section 4.2 Urbanization influences on basin runoff was rewrote, and Figure 6 was updated as another peak happening in flood season was added. It can be found, all three consecutive peaks during flood season were lower in urbanization scenario A than in current conditions. This phenomenon seems controversial at first glance. But the hydrological mechanism behind this was solid, as the description in last response. Because during flood season, the basin peaks were mainly contributed from large areas of rural and urban green areas. Therefore in urbanization A, faster urban sub-flow spread the peak over a longer period of time, hence reduce peaks in the total runoff.

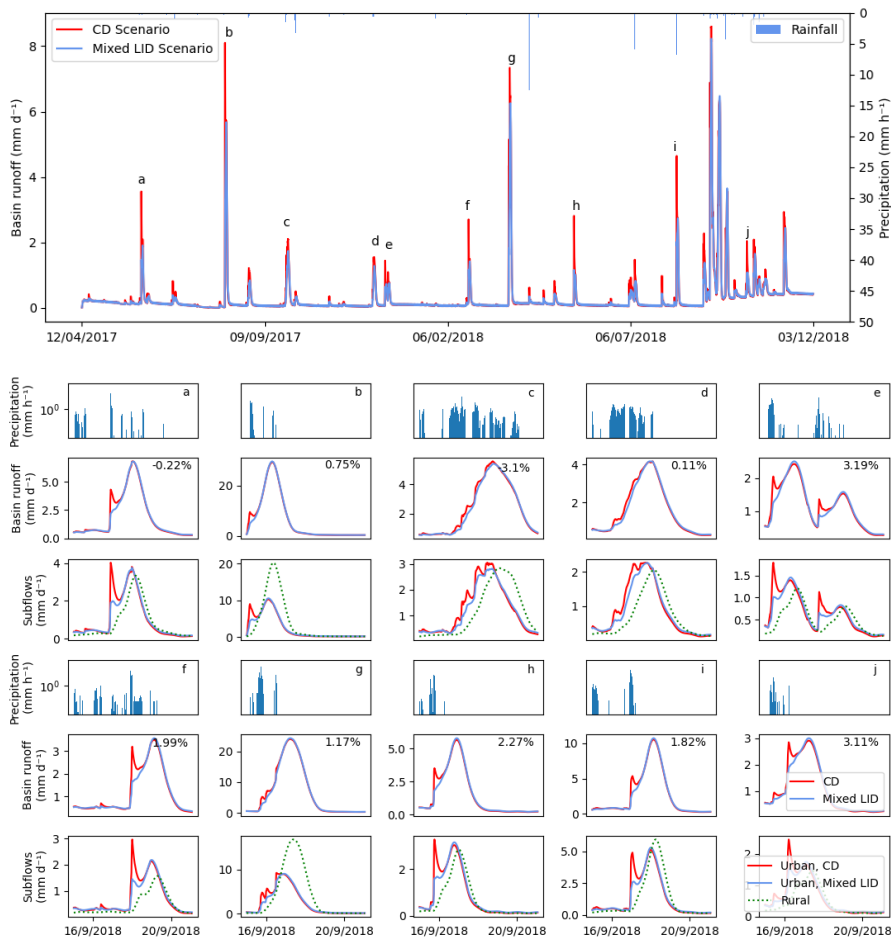
23) *A general weakness of this paper is that the results and the main findings are not discussed in the light of previous relevant studies in the results section, and most importantly in the discussion section. There are numerous studies on these topics. (page 20, line 435)*

Response: Thank you for this valuable advice. Some comparisons about the performance of four LID practices between literatures and this research results were added in Section 4.3.1-4.3.4. Besides, another Section 5.3 Comparative analysis was created in the Section 5 Discussion to compare some arguments in this research refer to former studies.

24) *The reduction of peak reduction ability of LID in wet periods and for big storms is logical and was also discussed in previous studies. However, The coincidence of rural and LID peaks that leads to a small increase of the combined peak can be the result of a coincidence. Considering also the uncertainty of the model performance as regards LID hydrological response, I believe that more caution is required before drawing these conclusions. (page 18, line 463)*

25) *Please see my previous comment. I believe that the obtained results cannot justify this as a general conclusion for combined urban and rural watersheds, considering also the modeling uncertainties. This could be the case in very specific cases, but not generally. (page 23, line 496)*

Response to 24 and 25: We appreciate this suggestion. Although completely avoid of model uncertainty is impossible, more tests about the rainfall-runoff relationship are helpful. To further

confirm LID influences in flood season, ten precipitation events with different rain intensities and durations were selected from the 600-day precipitation observation. These selected rain events were tested after 15th Sep. 2018 (flood season with the saturated subsurface soils in rural areas) in mixed LID scenario. Ten original peak runoffs corresponding to the selected rain events and ten test results were shown below.

Among ten test precipitation events, eight basin peak values were increased from 0.1 % to 3.2 % after the implementation of mixed LID practices, while only two peaks (a and c) were decreased by -0.2 % and 3.1 %. Especially for three extreme large rainstorms, the peak values increased by 0.75 % (b), 1.17% (g), and 1.82% (i). Even though the increase is small, it is to be concluded that during extremely wet conditions, the effect of implementing LID measures on peak flow reduction is negligible, if not negative in basins with combined urban and rural land use.

Based on this test, a paragraph was added in Section 4.4 LID performance in flood season. Considering the limitation of space, these two figures were not shown in the manuscript.

26) *As you mentioned in the previous sentence, this could be the result of a specific coincidence. By extending this conclusion, the natural watershed should present higher peaks than the urbanized one. In order to be able to justify this conclusion the model should be calibrated and validated separately for the urban and the rural parts of the watershed (this could be possible; as I can see in figure 1 there are hydrometric stations at the outlets of the urban and the rural watersheds). The LID modeling performance should be also justified, as any differences in the lag and the retention may change the obtained results. (page 23, line 500)*

Response: Thanks for this comment. As you expect, the runoff data collected from the urban and rural sub-catchments were used. As mentioned in Section 3.2.1 Semi-distributed model setup, the hydrological model starts from two simple lumped pre-models, one for a rural and one for an urban sub-catchment, respectively. In this process, the data collected from two sub-catchments (rural and urban) were used to calibrate the two lumped pre-models, respectively. Then, the dominant water processes was identified from lumped models and inherited by semi-distributed models for the simulation of the whole study catchment. During the selection of semi-distributed model, the runoff characteristics of simulated urban and rural sub-flows were also compared with the runoff timeseries of two sub-catchments.

For the LID model module, the parameter uncertainty was admitted and discussed in the section 5.2 Limitations as "LID implementation scenarios presume optimistic LID implementation conditions by using favorable LID parameters, hence overlooking practical implementation, operation, and maintenance problems such as the damage of LID practices and the blockage in soil media."

Besides, since it is unrealistic to avoid model uncertainty completely, different LID implementation scenarios with various types of LID practices and different construction extents were designed to provide results as reliable as possible. According to the model results, for all the 5 LID scenarios and 2 precipitation events, 9 of 10 basin peaks were increased after the implementation of LID. This is admitted that the specific increase numbers may be fluctuated due to parameter uncertainty. But the risk of increasing basin flood can be proved. And also, the hydrological mechanism behind this was solid as the LID practices delay the urban sub-flows and cause more overlap of rural and urban peaks, which increase the basin peaks in the end.

27) *A key limitation is that the model was calibrated and validated only for the current conditions and not for LID. The model structure is different for the LID scenarios and new parameters are involved. Therefore, the ability of the model to effectively describe LID is questionable. I believe that a solution would be to use the same model structure in all scenarios and change some parameters corresponding to the areas covered by LID or something. As I also mentioned in the relative section, it isn't clear how the parameters values related to LID were obtained. (page 24, line 513)*

Response: For the selection of LID parameters, it was answered by response 13. For the calibration of LID model module, as the response 17, since the LID implementation scenarios are designed, there is no measured data for calibration. In addition, for this kind of ungauged basins, that parameter calibration can be replaced by estimating with prior information has been proved by some studies (Sivapalan et al., 2003; Breuer et al., 2003; Hrachowitz et al., 2014; Gharari et al., 2014).

Breuer, L., Eckhardt, K., and Frede, H. G.: Plant parameter values for models in temperate climates, Ecological Modelling, 169(2-3), 237-293., 2003

Gharari, S., Hrachowitz, M., Fenicia, F., Gao, H., and Savenije, H. H.: Using expert knowledge to increase realism in environmental system models can dramatically reduce the need for calibration, Hydrology and Earth System Sciences, 18(12), 4839-4859, doi:10.5194/hess-18-4839-2014, 2014

Hrachowitz, M., Fovet, O., Ruiz, L., Euser, T., Gharari, S., Nijzink, R., Freer, J., Savenije, H. H., and Gascuel-Odoux, C.: Process consistency in models: The importance of system signatures, expert

knowledge, and process complexity, Water Resources Research, 50(9), 7445-7469, doi:10.1002/2014wr015484, 2014

Sivapalan, M., Takeuchi, K., Franks, S. W., Gupta, V. K., Karambiri, H., Lakshmi, V., . . . Zehe, E. (2003). IAHS Decade on Predictions in Ungauged Basins (PUB), 2003–2012: Shaping an exciting future for the hydrological sciences. Hydrological Sciences Journal, 48(6), 857-880. doi:10.1623/hysj.48.6.857.51421

28) *However, the key question is what is the probability of this to happen? Is it a rare coincidence or a typical outcome for big storm events in the case of a large urban area located at the downstream part of a rural watershed? (This question was extracted from the review letter.)*

Response: We thank you for this valuable question. The expression about the specific geographical condition sounds limited. However, it is quite common and may happen at every river segment downstream in the city. For example, if we selected one random river section at the urban downstream as one catchment outlet. In this catchment, the urban areas will be closer to the basin outlet than other rural areas.

Yours sincerely,

Xinxin Sui and Frans van de Ven