Q1: The forecast effects of the LID measures are not validated against any measured data.

R1: Our paper does not deal with a unique LID/NBS measure, but gives insights on how to optimise the choice among a set of them. Obviously, it is impossible to have the means to empirically validate a set of measures on a given basin, but this does not prevent to partially validate the numerical simulations with a degree of confidence that is closely related to that of the validation of the baseline scenario. In the present case, the latter was achieved with a high degree of confidence (e.g. the NSE coefficients for the event 16/09/2015 is 0.99)

Q2: "In the whole catchment, each LID/NBS scenario is more effective in two stronger but short events." I do not understand what is meant here.

R2: The two stronger events are composed by several short but intense sub-events and some inter dry periods. The sentence means the LID/NBS practices have better performance in terms of the intermittent rainfall (sometimes intense). To our knowledge, this is a characteristic that has been largely unnoticed so far.

Q3: It seems obvious to me that different sub-catchments with different characteristics will respond differently to the implementation of LID measures. The effect of representation of spatial variability of land cover and precipitation is not actually tested in the paper.

The first sentence supports the main goal of our paper and explains its main mechanism. But, contrary to the second sentence, our paper did deal with it. As discussed above, what could have been missing is a direct comparison with low spatial variability modelling that we are willing to introduce in the revised version.

Q4: L72-73: If some research does use more detailed data (as this sentence states) then this is the most relevant literature to be reviewed in the introduction, yet no references are given!

L72-73: In general, rare research used the high resolution rainfall data (i.e., X-band radar data with 250 m spatial resolution) and considered the coupling effect of the spatial distributions of precipitation and land uses.
We have to confess that the word "rare" is too diplomatic and confusing. This sentence will be corrected to clearly state that to the knowledge of authors, this is the first time that X-band radar data are used as meteorological inputs to analyse the hydrological impacts of LID measures, and to consider the coupling effect of the spatial distributions of precipitation and land use.

Q5: L127: The resolution of the DEM is coarser than that of the model. This limits the value of having the model at that resolution, and it may also lead to problems with the surface runoff module. Is there no higher resolution DEM available?

In this study, the available DEM data had a 25 m resolution and they were downscaled by interpolation to the resolution of 10 m. We agree that the results could be more reliable with higher DEM resolution will enable to run Multi-Hydro with higher resolution as well, but these data are not presently available.

Q6: L33-138: Using only three sampling points for soil classification may be too limited. Although it may be appropriate for the deeper soil layers, studies have shown that urban areas have a high degree of spatial variability in the top layer of the soil and/or the infiltration capacity. Combining a fully-distributed model with uniform data runs the risk of getting the worst of two worlds, i.e. lots of work to set up the model, but not actually using more information than coarser modelling approaches.

L133-138: Soil data for the catchment was obtained from the InfroTerre Database (http://infoterre.brgm.fr). As shown in Fig. 3, the soil data of three points are selected, which indicates that sand clay is located at the first layer of the soil profile. For point 1, there
135 exists a layer of limestone soil, which is less permeable. Silver sand which has the best infiltration ability, was in the next layer. The last layer is silver sand and clay, which has better infiltration ability than limestone. From a hydrological point of view, the soil data shows the complexity of the subsurface of the catchment. Therefore, the soil profile is reasonably simplified into three layers: sand clay layer (0-10.5 m), silver sand layer (10.5-25.8 m) and silver sand and clay layer (25.8-40 m).

In Multi-Hydro, the infiltration process is a two-stage process:
-firstly estimated with the help of the surface module based on the highly variable surface data (land use)
-secondly, if this first estimate is strong enough, this infiltration in the soil is estimated with the help of the infiltration module with possible feedbacks to the previous step estimate.
We certainly have to explain that although we are interested by the highest available resolution of all data, those of the soil have usually a much coarser resolution than the surface data, but fortunately have only a sort of second order impact. This explains why commercial softwares drastically reduced the number of available layers with respect to their initial versions.

Q7: L163-L178: The proposed scenarios assume that LID measures are implemented on all suitable surfaces, is this realistic? L172-175: Although steep roofs are typically unsuitable for green roofs, green roofs may have gentler slopes.

L163-L178 : For rain garden scenario (Fig. 6b), the low elevation greenbelts around houses were implemented by rain gardens, which can collect and store up the surface runoff from surrounding impermeable areas before infiltration on site. When rain garden saturated, the redundant surface runoff will drain into the drainage system. On the basis of application condition of rain gardens and the urban planning of the city of Guyancourt, 11.5 % of the whole area is set as rain gardens in the catchment. In the catchment, most of the buildings are houses with sloped roofs. Other types of buildings with flat roofs, only constitute one third of the total building area. According to the properties of green roof, small and light green roofs consisting of a soil layer and a storage layer are implemented on all flat ones, which can be simulated by the green roof module. All slope roofs remain unchanged. Finally, green roofs were applied to 11.5 % of the whole area (Fig. 6c). The combined scenario (see Fig. 6d), combined the three aforementioned LID/NBS practices. Those practices occupy 37.5 % of the whole catchment. In this case, the area of pervious surface reached

4.6 km2, which is about 88.4 % of the whole catchment.

Indeed, the proposed scenarios were based on a maximal implementation with respect to given physical properties of LID measures and site conditions. There is no difficulty to (randomly) select only a proportion of them to take into account other factors such as costs, social acceptance etc., as well as to modify site conditions, e.g. to enlarge the potential green roofs by including roofs with gentle slope.

Q8: L180-186: Evaluated only based on peak flow and total runoff volume, alternatively, are there other relevant metrics where the benefit of the detailed approach would be clearer?

L180-186: The general hydrological response of five scenarios under three rainfall events has been assessed by Multi-Hydro model and its green roof module (3×5 simulations). Two index (ΔV, total runoff volume reduction; ΔQp, peak discharge reduction) are calculated:

$$\Delta Qp(\%) = \frac{Qp_0 - Qp_i}{Qp_0} \times 100$$

$$\Delta V(\%) = \frac{V_0 - V_i}{V_0} \times 100$$

where Qp_0 *refers to peak discharge and* V_0 *represents total runoff volume in baseline scenario, and the* Qp_i *and* V_i *are peak discharge and total runoff volume of other scenarios with different LID/NBS.*

Certainly, there are other metrics to better highlight the space-time heterogeneity of the basin response and we are currently testing some of them to be included in the revised version, e.g. local ΔQp(%) and ΔV(%), as well as their quantiles. However, the peak flow and total runoff volume are presumably the most suitable metrics to make the comparison of the results of homogeneous and distributed data and modelling.

Q9: L204-206: There is a major deviation for the third event. Wouldn't it be better to calibrate the model further so that it better simulates the catchment behaviour?

L204-206 All the model parameters related to the land use type and soil type were selected from the Multi-Hydro model manual (Giangola-Murzyn et al., 2014), as shown in Table 2 and Table 3.Green roof is a special LID/NBS practice which needs to be simulated with the Multi-Hydro green roof module. The properties of the green roof are illustrated in Table 4.

Multi-Hydro being physically-based, does not require calibration, contrary to conceptual models and the expression "model parameters" is somewhat misleading. It should be replaced by physical constants (e.g. conductivity) defined by land use and soil types.

The major deviation of the third rainfall events could probably be due to that the highest rainfall peak of radar data only last 3 min, and the simulated retention pond couldn't accumulate enough water compare to the measurement data. We simulated the same rainfall event with the C-band radar data with the spatial resolution of 1 km, and temporal resolution of 5min. For the validation results of the C-band radar data, the major deviation is less obvious than that of the results of X-band radar (see additional Fig.8). This could be explained by the fact that C-band radar is closer to the catchment and could better detect the rainfall intensity. Our co-authors indicated (Paz.et al, 2018) that the X-band radar tends to underestimate the total accumulated rainfall depth.

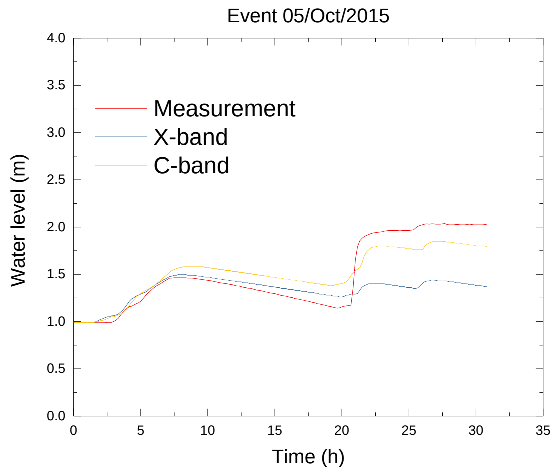

Fig.8 (additional) Comparison of observed and simulated water level with two different type of radar data (X-band and C-band) of event 05/10/2015

Q10: L212: A model result with a NSE close to 0 is not "credible", see e.g. Moriasi et al (2007). After all, 0 is the score that would be achieved by the average of the observations, which is not a particularly strict (albeit a commonly used) benchmark.

L213-214-The Nash is closer to 1, indicating that the model is more reliable. Nash is closer to 0, indicating that the simulation result is closer to the average observed value, which means the result is credible.

We apologise for the typo in L214: "which means the result is credible" should indeed read "which means the result is not credible"

Thanks to the helpful reference.

Q11: The scale of the y-axis in the graphs leaves a lot of empty space. It would be better to use this space to show the data in more detail, as it is now difficult to see differences between the different lines.

We modified the graph according to your suggestion. The modified graphs are attached.

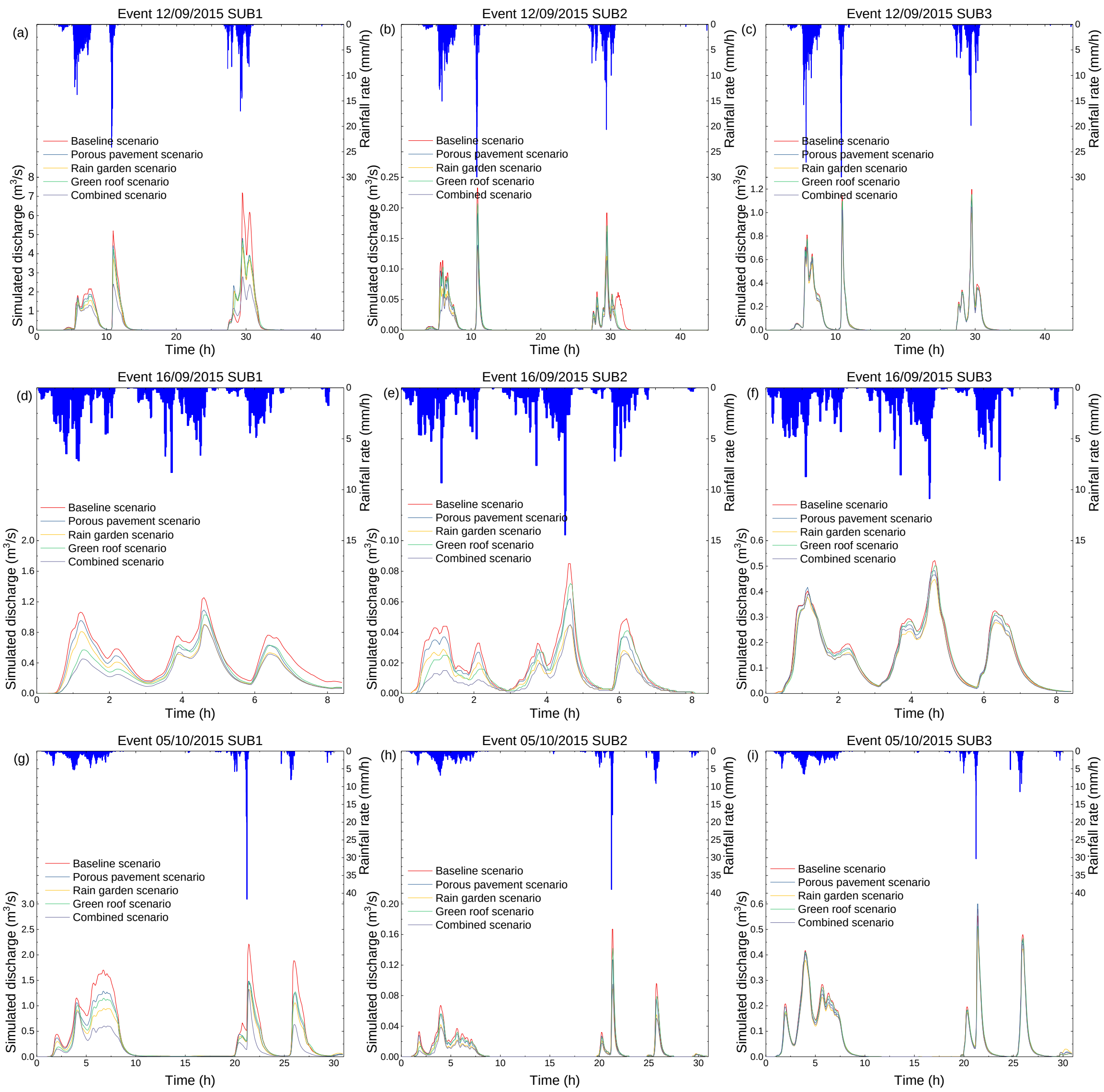

Figure 11(additional): Presentation of the simulated hydrographs (a to i) in three sub-catchments for three rainfall events and five scenarios.