Author's response to referee # 2

Our responses are in blue in between the original comments.

In this article, the authors describe the physical representation of green roofs in the Town Energy Balance (TEB) scheme. Although this is an important research subject, and although I found some aspects of the study quite interesting, this paper is seriously flawed and I cannot recomment it for publication in GMD.
The first problem concerns the lack of focus of the study. Indeed, it is not clear what exactly are the main objectives of this work (they are not stated in the Introduction) and what the authors wish to accomplish. Because of this vagueness in the authors' aim, it becomes quite difficult for the reader to pose any judgement on the results presented.
I understand that GMD mostly focuses on the description and evaluation of numerical models used in geosciences.

We are not sure what the referee means by "lack of focus of this study". We believe that the objectives of our study, the modelling of green roofs within the TEB model to quantify their benefits (in terms of building energetics, thermal comfort and runoff, mentioned at local scale in the Introduction paragraph) at the scale of cities are quite clear (abstract lines 15-18 & p 3 lines 6-9 & p 4 lines 3-4). The revised manuscript aims to provide a better description of the model (see paragraph 3.2) and the evaluation of the model for a case study is done in two stages. The first stage consists in a hydrological calibration (paragraphs 4.4 & 4.5), which objective is to determine the type of soil (organic matter, sand or clay) that best describes the hydrological behaviour of the plot drainage layer (as those are not well known). The second stage is the evaluation of the calibrated model over the remaining time period (paragraph 4.6). Please see revised manuscript.

But my opinion is that the approach presented by the authors is not evaluated properly, and that the presented conclusions are in error. My greatest concern is related with the "calibration" aspect of this work. In their quest to obtain the best solution, the authors performed a total of 576 runs based on all possible combinations of physical parameters for the vegetation and soil / substrate layers on the roof. As shown in Fig. 4, these runs are of a widely variable quality when evaluated against one of the chosen variable (here soil moisture). The choice of a single solution, or "calibration" as stated in the text, suggests that these best results could have been obtained by pure chance, and does not demonstrate the quality of the physics that was implemented in TEB. Considering that the results significantly depend on the specification of the parameters, it might be argued that a simpler representation of the physical process could be preferable. The tests performed by the authors look like an ensemble experiment. It would have been more interesting, I think, to examine the performance of all the members (or of the mean), and examine the uncertainty of the simulations as expressed by the spread.

Neither the vegetation characteristics nor the soil thermal properties did change in the initial 576 simulations, only the hydrological characteristics did. We agree that running the model with all the possible combinations of hydrological characteristics meant that some of them did not have a physical reality, which explained the "widely variable quality of the runs" mentioned by referee # 2. To overcome this, the calibration stage was rethought based on three ensembles of runs whose characteristics have a physical reality. Since the hydrological behaviour of the materials constituting the drainage layer of the green roof studied are not well know, each ensemble describes a typical hydrological behaviour (that of organic matter or sand or clay) for this layer (paragraph 4.4). What then creates the ensemble elements is the variation in the other hydrological parameters which

originates from multiple data sources. The number of runs examined was consequently reduced to 3x32 (p 17 lines 13-19).
The performance of the model for the calibration of the drainage layer was then examined in terms of ensembles (means and min, max) by using scores and graphs (Figures 4-5-6), which clearly highlighted a best ensemble(paragraph 4.5). We believe that this new type of calibration exercise does show that the best ensemble is not “obtained by pure chance”.
This best ensemble was then evaluated over a different time period (paragraph 4.6), using the same statistical scores and a better subset was identified, providing us additional information on the in-situ porosity of the substrate (p 21 lines 14-16).

Even if we accept the authors approach, then we are left with a best solution which does represent that well the observations. Results for the soil moisture are reasonably good, but there seems to be some problems for the drainage. And the substrate temperature has serious issues (bias and amplitude of the diurnal cycle), in spite of the authors’ claim of success.

Firstly, in order to rerun the analysis, the simulations were re-run with a slightly updated version of the model and the vertical discretization of the soil layers used for calculations was modified to improve numerical stability (from 18 to 6 sub-layers) - p 16 lines 6-9.
Secondly, the choice of the calibration period was chosen carefully (from p 17 line 20 to p 18 line 2) to take into account as many physical processes as possible and to avoid time periods at which the recorded outlet drainage might be biased (in relation to the stoppers installed at the green roof base aimed at reducing the loss of water for, p 17 lines 23-30).
In the initial manuscript, the entire time series were analysed and errors partially explained by these stoppers ; this was not easy. We found it clearer and fairer to GREENROOF to explain the device of stoppers, which the model does not model, and to re-run the calibration at a time period when the records of drainage were the least disrupted.
Ultimately, the scores for the calibration are much better over the new calibration period be it for the substrate water content or the outlet drainage. To illustrate this, a Taylor diagram for the outlet drainage has been added to the initial manuscript (Figure 4) and calibration scores have been gathered in small tables at the bottom of new figures (Figures 5 and 6).

Following the calibration, GREENROOF was evaluated based on the best ensemble performance. Within this ensemble two subsets with different hydrological behaviours were identified (p 20 lines 1-18 & p 21 lines 10-16) and the evaluation scores are therefore presented for the best ensemble (and its extremes) and the two subsets in terms of the model hydrological and thermal performance.
The effect of water content on the positive bias and the higher amplitude of simulated temperatures has been investigated and is added to the manuscript (from p 20 line 32 to p 21 line). As it only contributes partially to the bias, additional hypothesis are mentioned but can not be evaluated because a comprehensive set of data, including the surface fluxes (LE, H, G) would be necessary to study these aspects, and they are not available at the case study site.
However because this was maybe vague for scientists who do not know ISBA (this was mentioned by referee # 1), it is important to highlight that the estimation of surface fluxes is realized in ISBA by a detailed parameterization (that of plant transpiration is presented in the new manuscript paragraph 3.2.1 p 8-9) and that the heat storage flux at the soil surface is calculated based on the closing of the surface energy balance. While the analysis did not highlight any problem in the partitioning and order of magnitude of H and LE at first sight, the respective fluxes could only be checked with measurements. This is mentioned as a perspective of this work in the Conclusion p 22 lines 7-10.
We think that this case study highlights how difficult the evaluation of green roof models is, especially considering the difficulty to find case studies with complete datasets which allow the description of all the physical processes involved to be evaluated.

Finally, there are no demonstration that this set of parameters could be generalized to the city scale, as stated by the authors in the conclusion.
There are other minor aspects that would have to be corrected if this article is accepted for publication, related to the English language used (the paper is easy to read, but some sentences are awkward) and organization (many paragraphs are way too long and could be split in two, three, or even four shorter ones).

The way the analysis of our case study simulations was done in the revised manuscript (analysis of simulation ensemble, paragraph 4) does not retain a specific set of parameters anymore. It focuses more on studying the model response to an ensemble of hydrological properties rather than finding the best set of properties, even if one can draw conclusions from the calibration (paragraph 4.5, especially p 19 lines 12-19) and the evaluation (p 20 lines 10-11 & p 21 lines 14-16) exercises.
Also, the experimental plot chosen for the evaluation of our model is of a type that is very representative of the green roofs that are currently implemented in cities, as stated in the manuscript (p 14 lines 14-28 and p 22 lines 11-12). Hence the possibility to generalize the green roofs characteristics for modelling green roof benefits at the scale of a city, as stated in the Conclusion paragraph. The purpose of our modelling studies is not to model the existing green roofs implemented in a city but rather the impact of such a construction on the energetics of building, etc. With this in mind, we find it perfectly acceptable to generalize the soil characteristics of this case study at the city scale.
As far as minor aspects are concerned, as we re-wrote big parts of the manuscript in order to present the re-analysis of our case study simulations, we made the effort to split long paragraphs into smaller ones and to use the best English we could.