To whom it may concern:

We have completed the requested edits to our manuscript (egusphere-2026-22), Incorporating spatial heterogeneity into evapotranspiration estimates for bioretention basins. We are grateful to the reviewers for providing thoughtful comments, and to the handling editor for their work in finding reviewers and evaluating the work. We also appreciate your patience. We were not clear that we needed to finalize the review after making comments, which led to a delay. Also, several reviewer comments were ultimately quite helpful, but did lead to new analyses (in particular, see Figs S2-S4 and related text) that took us some time to complete. However, we have been extremely careful in addressing all comments and making other minor improvements to the figures, tables, and text. We look forward to seeing this work published in HESS.

Josh Caplan et al.

Reviewer 1

General comments

Very interesting article on how estimate at high spatial resolution the evapotranspiration fluxes (ET) in the very heterogeneous urban canopy. The subject is important for the estimation of the runoff reduction in green space but also of the water need of urban green space.

The article is outstanding due to the diversity of estimation, with measurements (chamber as the reference but also soil water content) and modeling (empirical-statistical as the reference but also a wide range of conventionnal formula).

The work is very clearly presented, with a very complete and rigourus methodology combining different aspects (shading, vegetation characterisitcs, statistical test, ...). The database with the characterisation of the site (soil, vegetation, topography, urban scene) and with the temporal series of different hydrological and meteorological variables is really interesting and outstanding.

The form is perfect, with clear text and illustration; I do not have a look to the data (https://www.hydroshare.org/resource/ba487d9c6e4b473f88d8298499ee1c6d/), due to a lack of time.

It is extremely gratifying to read these comments after working hard on this project for several years to make it as strong as possible.

Specific comments:

See the pdf file for specific and detailed comments.

I suggest "Minor revision" (rather than "accepted subject to technical corrections") because I suggest to add some details on the development & performance of the empirical model (adding a chapter in the Supplementary Information at least). The evaluation is presented positively in two lines in the article (L284-285), even though the model is subsequently used as a reference throughout the rest of the article. It would be useful, at the very least, to show and discuss the observed/simulated scattergram at the basin scale, and, if possible, at the various measurement points as well.

We have expanded the description of the model's development in section 2.5. We have also added further information on the model's performance in section 3.1, and have moved the table containing model coefficients and associated statistics to the main document (the new Table 3). Further, a new Fig. S2 was added, which shows key model diagnostics (residual normality and heteroscedasticity) and observed ET vs. modeled ET. There are two versions of modeled ET, however: one showing estimates when only fixed-effects are used and the other showing estimates when both fixed and random effects are used. The figures were made at the plot scale though, since that is the scale at which field measurements were made, and the scale of the data from which the model was developed.

Three other points could be add in the discussions part:

- the spatial role of other micrometeorological conditions than the radiation, for example the wind or the VPD? Do you have some site informations of these variability (contrast in the two weather station for example)? For another work, it could be possible / interesting to introduce such spatial variables in the empirical model?
  We agree that it would be ideal to incorporate spatial variation in more factors than we did, and have added the following paragraph to the study limitations section to address this point. "*Third, although spatial variation in solar radiation was explicitly represented in our ET quantification, other microclimatic factors were assumed to be uniform. This is likely justified because air (and thus its temperature and humidity) tends to be well mixed at the basin scale, whereas solar radiation can be extremely variable over short distances due to shading. Nonetheless, VPD and wind speed could vary spatially within bioretention systems, particularly where*

surrounding structures and vegetation alter microclimate and airflow. Future studies could incorporate spatial measurements of these variables to determine whether accounting for this additional heterogeneity improves basin-scale ET estimates." Although there were two weather stations, one was elevated well above the basin, so we do not see a reliable means of comparing the two in this study.

- At different lines in the article, the validity of conventional models (=formulas) is debated in terms of whether or not they take water stress into account (for example, lines 427–429). Upon examining the soil water content at -5cm (Fig. 3d), I find that the values are not particularly low (and one would expect higher values at greater depths) and that it is not certain that ET could be limited by soil water availability. This hypothesis is consistent with the shape of the scatter plots between ET values simulated by the empirical model and the different formulas: for example, this relationship is highly linear for the PM formulas across all ranges of ET values (Fig7), including high values whereas for these high values one might expect more severe water stress to occur. In summary, I get the impression that ET water limited situations are rare in the database, and therefore that the reasons for the offset between the empirical model and the PM models must be sought elsewhere;

In response to this comment and a related one from Reviewer 2, we added an analysis evaluating whether soil-water limitation contributed to differences between conventional and empirical ET estimates (Sect. 2.8; Supplement 1; Fig. S4). We adjusted conventional ET estimates using an FAO-56-style water-stress coefficient (K_S) calculated from the soil moisture profile and soil water-retention characteristics, using three plausible effective rooting depths (20, 30, and 40 cm). The analysis showed that meaningful water limitation was infrequent during the model-comparison period. As now reported in Sect. 3.5: "*Across the range of effective rooting depths we evaluated (20–40 cm), mean daily water-stress coefficients (K_S) ranged from 0.969 to 0.987.*" Further, "*There were periods when K_S fell below 0.75, but these accounted for only 6%, <1%, and 0% of days when using 20-, 30-, and 40-cm rooting depths, respectively.*" Using a 30-cm rooting depth, incorporating soil-water limitation reduced the overestimation by the PM models by approximately 5–7%, primarily at the upper end of the ET range, but also reduced concordance between empirical ET and all six conventional models. We therefore revised Sect. 3.5 to conclude that "*although not accounting for soil-water limitation contributed slightly to the higher ET estimates from the PM models, it is unlikely to explain their systematic offset from empirical ET.*"

- The linearity of the relationships between the empirical and PM models, which consistently involve a simple offset, is truly impressive. A theoretical comparison of the two models could be useful for analyzing the following observation:

    - the empirical model uses the following explanatory variables: topography (via a 1/0 index), vegetation height (spatial variability only?), VPD (temporal variability only?), solar radiation (spatial and temporal variability), and soil moisture status (temporal variability at a single point); it is statistically calibrated on 7 x 11 daily ETs;
    - PM formulas primarily use the temporal variability of solar radiation, VPD, wind, and temperature, with a constant vegetation height.

    So, for example, the effect of wind is taken into account in PM (via its aerodynamic term), which is not the case in the empirical model. Could this be a source of overestimation of PM ET?

We have added a paragraph to section 3.5 to address this point. It reads: "*The strong linear relationships between our empirical ET estimates and those from PM models likely reflect their shared sensitivity to the atmospheric variables that control ET, particularly solar radiation and vapor pressure deficit. However, the two approaches use very different methods to account for surface and aerodynamic factors. The empirical model was parameterized directly from measurements within BASIN A and incorporated spatial variation in plant height, radiation, and topographic position explicitly, but other factors (like wind speed) implicitly. In contrast, the PM formulations account for vegetation and aerodynamic factors only via their inbuilt parameters but explicitly incorporate wind speed. Wind speed did appear to influence the discrepancy between PM and empirical ET estimates, as differences increased significantly with wind speed for both PM formulations ($p < 0.001$; Fig. S3). Further, while wind speed explained 39% and 12% of the variation in these differences for PM-ASCE and PM-FAO models, respectively, it explained ≤5% for the other conventional models, suggesting that the treatment of wind speed contributed to PM estimates being greater than the empirical estimates. Although wind speed was not an explicit predictor in the empirical model, airflow generated by the chamber mixing fans influenced the measured fluxes and was therefore implicitly represented in model calibration. However, the magnitude of this influence is uncertain because wind speed within chambers was not measured. Moreover, wind speed data used in the PM models may not have fully represented conditions within the basin, since surrounding and internal structures could have reduced and/or accelerated airflow.*"

Comments from PDF

Line 54: One must always exercise caution when presenting results of this kind, as they combine extremely different contexts (particularly in terms of climate and associated runoff). See also for bioretention cells: Huang, T., Sage, J., Técher, D., Gromaire, M.-C., 2025. Hydrological performance of bioretention in field experiments and models: A review from the perspective of design characteristics and local contexts. Science of The Total Environment 965, 178684. https://doi.org/10.1016/j.scitotenv.2025.178684.

We have largely rewritten the paragraph to incorporate the caveat mentioned as well as the results in the review plus another paper. We believe that it is a much stronger paragraph now. It reads: "*The overall importance of ET to the removal of stormwater in bioretention basins is highly variable. In their review of ET in GSI, Ebrahimian et al. (2019) found that studies reported ET removing 19-84% of influent stormwater, though a modeling-based analysis estimated ET below 5% (Brown and Hunt, 2011), widening the range further. More recently, Huang et al. (2025) reported stormwater removal via ET ranging from near zero to almost 90% across bioretention systems. Although this broad range encompasses systems operating under widely differing conditions, some of the variation can be attributed to well-established effects of climate, hydrology, vegetation, and system design. For example, removal via ET tends to decline with increasing hydraulic loading ratio (Huang et al., 2025) and also varies with vegetation and media characteristics (Nasrollahpour et al., 2022). However, the factors underlying this variation are often inadequately represented in ET estimates, particularly when heterogeneity occurs within individual bioretention systems.*"

Line 116 (map): you can add also the location of the substrate water content sensor

The location of all soil moisture sensors is now indicated on the map. Because they are close together, they appear as a single symbol.

Line 120: Is it possible to add a photo of the chamber in the Supplementary information? Ideally taken during the experiment

We do have a photo, and it has been added to the supplement (the new Fig S1).

Line 227: could be located in Fig1a

The location of all soil moisture sensors is now indicated on the map. Because they are close together, they appear as a single symbol.

Line 229: Could you justify why you choose the sensor at -5cm in the empirical model and the other (deeper) for the soil moisture-based estimation?

We now introduce both soil moisture datasets in a single paragraph within section 2.3, Environmental Data, and explain why the uses differed: "*Volumetric soil water content (VWC) was measured using TEROS 12 sensors (Meter Environment, Pullman, WA, USA) installed on the south flank of BASIN A, approximately 40 cm above the basin floor (Fig. 1a). The sensors were a subset of those used in a separate study (Pope et al., 2025) and were within 1 m of each other. Sensors were positioned at 5, 10, 30, and 60 cm depth, with the latter three arranged in a vertical profile. For empirical modeling of ET (see Sect. 2.5), we used the mean VWC from 5- and 10-cm depths since these measurements most directly depicted near-surface moisture conditions. Missing values in the 10-cm time series were first interpolated from concurrent measurements recorded by nearby sensors, yielding complete data for 137 days (8 June through 22 October). The sensors at 10, 30, and 60 cm provided a soil moisture time series spanning the soil profile and were used in analyses requiring data for multiple depths (see Sects. 2.7 and 2.8). Profile data were missing for three of the days included in the empirical-model time series and could not be accurately interpolated (n = 134 days).*"

Line 240: "daily ET" if I understand? Please specify

Correct; we have updated the text accordingly.

Line 284: I recommend to show the performances of the statistical/empirical model (for example in the supplementary data, with a scatter plot between daily ET observed / simulated (for the all dates & plots)

As mentioned above, this has been added to the supplement (the new Fig. S2).

Line 297: yes, and also that the ET process is certainly energy-driven for a lot of the simulated days; the size of the vegetation does not provide any more energy!

Very good point; we have changed to sentence as follows: "*The fact that this effect was not larger was likely due to available energy constraining the extent to which greater LAI could increase ET, as well as larger plants shading the soil surface more completely, inducing a tradeoff between soil evaporation and transpiration.*"

Line 302 (fig 2): It is not obvious to make the connection between the legend of the figure and the location in Table I.
Reviewer 2 also commented on this. We have updated Table 1 to use the terms from Figure 2, namely large, small, and minimal.

Line 317: I am wondering how much confidence we can place in the model for very tall vegetation, given that the measurements used to calibrate the model were taken on smaller plants (if I have understood correctly). Could you comment on this rather significant and uncertain extrapolation?

This is an important point, but one we address in the Study Limitations section (now 3.10). We suspect that the reviewer did not yet read that section when this comment and the next were written. We think the matter is addressed adequately but can expand if requested.

Line 324: idem, see previous comment

See above

Line 345: I completely agree, and that's a very interesting observation regarding our ability to measure ET in situ on urban green space (see also Ouédraogo, A.A., Berthier, E., Ramier, D., Tan, Y., Gromaire, M.-C., 2023. Quantifying evapotranspiration fluxes on green roofs: A comparative analysis of observational methods. Science of The Total Environment 902, 166135. https://doi.org/10.1016/j.scitotenv.2023.166135)

We appreciate the reviewer bringing this paper to our attention. We have incorporated its findings as follows: "*Although the soil moisture-based method has been validated using weighing lysimeters set up to mimic stormwater cells (Hess et al., 2021), comparisons among observational methods have demonstrated that ET quantification in GSI systems has substantial uncertainty (Ouédraogo et al., 2023). It is therefore not possible to definitively determine which approach more accurately reflected true, basin-scale ET.*"

Line 409: didn't understand. In reality, the relation PET/AET is certainly non-linear

We have reworded the last sentence of the paragraph. It now reads "*This approach was particularly effective for models whose ET_0 estimates had approximately linear relationships with our empirically-determined ET values, such as PM-ASCE and PM-FAO.*"

Line 436: I do not understand how realize this type of measurement campaign for the design of a basin (the basin does not exist yet). I'm also curious to know how the ET could be taken into account in the design procedure?

We have updated the paragraph to clarify that we meant that correction factors could be determined from existing basins in a region. See section 3.9.

Line 464: I totally agree, and for example I will be very cautions on the very high value of cumulative ET in Fig4b (>600mm in some segments with high vegetation)

Very glad that we agree.

Line 470: Estimate the water budget of the basin with different measurements could be very interesting (and so also to considered the deep infiltration). Your very rich database on the basin characteristics would be a huge asset for that work

We appreciate this comment and would like to do this (if time allows) at some point in the future.

Reviewer 2

I welcome the authors efforts to address this complex challenge. The work provides valuable real insights into the heterogeneity encountered when attempting to characterise bioretention basins. The chamber data and spatial characterisation of the basin are valuable and useful contributions to science in their own right.

We greatly appreciate the reviewer's supportive comments!

I feel the paper should be published subject to the following comments being addressed:

The naming of the three approaches, e.g. in the abstract, does not seem to be as precise and unambiguous as it could be. (1) and (2) both refer to 'empirically-based estimates' (vague), and could perhaps be better distinguished by specific reference to (1) chamber measurements undertaken at multiple locations and (2) a single vertical profile of measured soil moisture content.

Good point; we have implemented the suggested names.

The development of the statistical model (1) – Section 2.5 – is difficult to follow and not particularly transparent. The parameterized model, i.e. the final selected regression coefficients and parameters, should be clearly presented in the main body of the paper (e.g. within the Results, Section 3.1).

We have attempted to improve the explanation of the model's development. Readers unfamiliar with mixed-effects linear models may still struggle to understand the technical details, but such readers can easily find other sources for the information if desired. We believe the current version strikes a reasonable balance. Also, the table has been moved into the main document (Table S2 is now Table 3).

Define R^2_{marg} and R^2_{cond}.

This was done in the methods section but may not have been sufficiently clear; the explanation has been improved. Also, we have shortened the subscripts to m and c.

For completeness, (for readers unfamiliar with the use of chamber experiments) please state how ET in mmol m^{-2} s^{-1} is converted into ET in mm day^{-1}. This could be included in the Supplementary Material.

We have updated the last paragraph in section 2.5 to explain this, as well as to better explain how segment-specific values were aggregated: "*Daily ET for each segment was multiplied by segment area, summed across all segments, and divided by total basin area to obtain daily, basin-scale ET from 8 June to 22 October (n = 137 days). Basin-scale ET was converted from mmol m^{-2} d^{-1} to mm d^{-1} using the molar mass and density of water (18 g mol^{-1} and 1 g cm^{-3}, respectively).*" The supplement has also been updated to better explain the calculation of ET.

I can see the value of comparing ET estimates based on soil moisture measurements against the chamber experiments, but I don't understand the logic of deliberately trying to capture the heterogeneity with the chamber experiments compared with sampling at only one location with the soil moisture sensor. This is not really comparing the two measurement techniques, rather it's comparing two spatial sampling approaches while also varying the measurement technique! Why did you not use multiple soil moisture sensors to better understand the localised variations in ET? This also leads to some speculation around line 346-7. The authors have noted that multiple sensors would be better (Line 350-351), but this doesn't remove the question around experimental design. Consider simply introducing the soil moisture measurement approach as an order-of-magnitude sense check to the chamber-based model.

This is a good point and agree that additional justification for the approach is warranted. We have done this by adding the following sentence: "*This is a relatively simple method previously developed for estimating ET in bioretention systems (Hess et al., 2021), though it does not account for spatial variability unless used at multiple locations.*"

I am confused about whether/why moisture stress factors were, or were not, included in the application of the conventional models, Section 2.8. For FAO PM, normal practice would be to apply a water stress factor to the reference ETo in order to estimate actual ET. However, the commentary in Section 3.5 suggests that you did not include water stress. It looks as though (Supplement 3) you (potentially) included soil moisture content in the development of your empirical model, but then excluded it from your comparisons against the conventional ET models – I'm not sure that this is a fair comparison.

We agree that this required further evaluation. As described in our response to Reviewer 1 above, we added an analysis explicitly accounting for soil-water limitation (Sect. 2.8; Supplement 1; Fig. S4). The analysis showed that water limitation was infrequent and explained only a modest portion of the discrepancy between conventional and empirical ET estimates.

I understand that you have used K_L as a calibration coefficient to scale the estimates from the conventional ET models to better match your chamber measurements. I believe that the values ranged between 0.5-1.0 (Line 393). Please refer to Table 3 at this point in the text, as the actual values are easy to miss within the table.

We have added a reference to what is now Table 4.

The additive correction does not seem to be well justified or transferable, and could potentially lead to negative ET estimates. Is there a precedent in the literature for doing this? This is likely to be quite case-specific and not generally transferable. The reader can already see the offsets in Figure 7. It would be more useful to explore whether the inclusion of moisture stress factor addressed some or all of the overprediction observed with the conventional models.

This is another good point, and we have addressed it by adding contextual information and stronger cautions to the relevant sections. We also cite a precedent for using additive correction. See above for how we addressed moisture stress. Here are the main edits:

Methods: "*We also tested additive correction, a simple approach that can be used to adjust for model bias when it is consistent across a sufficiently wide range of values. It has been applied to estimates of various meteorological parameters, including ET (Paredes et al., 2018), though it has the potential to yield negative values when the correction exceeds the uncorrected estimate. Such values would have little influence on estimates aggregated over longer periods as long as they are small and infrequent. To determine the correction factors, we first regressed each set of* ET_{ref} *estimates against* ET_{emp} *and then subtracted the fitted intercept from each* ET_{ref} *estimate when the intercept was positive. As with the landscape coefficient adjustments, we used the CCC to assess how additive corrections improved* ET_{ref}*.*"

Results & Discussion: "*Additive corrections were more effective than landscape coefficients since these three models systematically overpredicted ET relative to our empirical approach. However, additive correction also produced negative daily ET estimates when uncorrected values were small, highlighting a limitation of this approach. Nonetheless, for these three models, additive correction yielded the greatest agreement with our empirical estimates.*"

Conclusions: "*Depending on the accuracy required, these model estimates could be used without correction, with the understanding that they are likely to overestimate ET, or they could be adjusted using additive, multiplicative, or shade-based corrections to improve agreement with empirical ET. We note that additive correction is highly context-specific and* *may produce negative estimates at low ET rates and should therefore be used with caution."*

Lines 464-466. Again a request to actually share the fitted statistical model so that the reader can see for themselves how ET varies explicitly with plant height.

The table has been moved to the main document and a reference to it has been added to the paragraph mentioned (in the section on study limitations).

Minor corrections:

- Line 25 – 'align with ours' should be 'align with the ET estimate based on the chamber measurements'.

  This has been corrected.

- Line 67 – 'such multiplying' should be 'such as multiplying'.

  Thank you; corrected.

- Line 106-108 – where exactly is the 'rocky gabion'? What do you mean by 'above', 'below' and 'lower' in this context? Are these relative elevations, or upstream/downstream or north/south? Please also check lines 170-171 re the use of the terms lower and upper.

  We have rewritten the sentence on the gabion to clarify: "The basin is approximately linear (148 m in length and 5 m in width) and has a trapezoidal cross section (~1.5 m depth). The middle 44 m of the basin's length contains a rock-filled gabion that causes the basin sections at either end to remain hydrologically separated except during extremely large storms. For the purposes of this study, we focused only on the portion of the basin to the southwest of the gabion (Fig. 1a)."

  We believe this rewording describes the layout better and avoids confusion with references to topographic position (upper/lower).

- Line 247 – should 'raw time of solar radiation' be 'raw time series of solar radiation'. I don't think you adjusted the time.

  Yes, this should have said 'time series.' Thank you for catching this.

- Do 'tall' and 'short' (Table 1) equate to 'large' and 'small' (Figure 2)? Please use consistent terminology.

  They do indeed refer to the same categories. We've updated the Table and associated text to match the figure.