Reviewer 1

The article presents a large dataset of 25 ecosystem structure metrics for the entire area of the Netherlands at 10 m resolution. The metrics are calculated from openly available airborne laser scanning (ALS) data and across multiple ALS campaigns, making it a highly valuable dataset to assess ecosystem dynamics. The article is well-written, with great attention to detail – I particularly commend the tables and figures that present a lot of information without feeling overly complex. The paper also follows a nice logical flow, from presenting the ALS pipeline, to detailed justifications for the derived metrics, to 2 sample case studies that have high relevance for applied research (changes in ecosystem structure + a comparison of structural indices across different ecosystem types). The resulting product provides insight into high resolution ecosystem change over at least 15 years from 2007-2022 (Note that I am not counting the first ALS campaign which likely does not reach minimum quality standards for ecological analysis). It thus has great potential to become a standard tool both for practitioners and researchers interested in ecosystems in the Netherlands. It could also become a nice reference dataset for similar efforts at larger scale.

Response: Thank you very much for your positive evaluation. We appreciate your valuable comments which have helped us to improve the quality of the manuscript. We have thoroughly addressed your comments one by one. See details below.

I will now provide my main comments, with line-by-line comments following below.

Main comment 1: Robustness of pipeline to pulse density / leaf phenology

The main issue where the authors have not (yet) convinced me is whether comparisons in time between different AHN surveys are robust to acquisition properties. Large disturbances (e.g. clear cutting, logging) will obviously be visible and can be separated from noise, but how about growth or smaller disturbances? The authors mention at the end of the introduction that intercomparisons between different instruments and scanning conditions could lead to considerable errors, but then do not really provide means to assess the sensitivity of the products or to correct for some of these problems. This is important, because the surveys of AHN1 and even AHN2 differ strongly from AHN3-5 in terms of pulse density, and even the more recent high quality surveys may differ in leaf phenology. E.g., a scan in April will likely already measure some vegetation in early leaf-on conditions, and this might create bias compared to a scan in December.

Overall, I see three points that I would like the authors to address:

1. **Sensitivity analysis:** The authors should provide a sensitivity analysis of the pipeline with regard to pulse density, e.g., how much do the inferred 25 metrics change when the pulse density of a high quality scan (AHN4-5) is degraded to levels of AHN1 or AHN2. This can then be used to provide clear bounds on what is ecologically interpretable. The simplest way to do

this would be to a) use the site in case study 1, degrade the point cloud from AHN5 and check robustness of the metrics, and b) to repeat case study 2, with the exact same sites and metrics, but with AHN1 or AHN2. Is the analysis reproducible for ecosystems with low vegetation and small differences between them? I do not expect all metrics to be perfectly stable for this study to be published, but having an estimate of uncertainty for all of them would be key. Note that I would degrade pulse density (the number of shots) and not point density (the number of shots + returns) to accurately simulate a lower-quality scan.

Response: Thank you for your insightful comment. We agree that the robustness of LiDAR metrics against varying pulse density is crucial for the interpretation and usage of the data products. We have therefore conducted a comprehensive sensitivity analysis on the robustness of the generated 25 LiDAR metrics based on different pulse densities. Note that for the four completed AHN surveys only the AHN3 and AHN4 provide pulse information (e.g. "return number", "number of returns") in the point cloud, whereas the AHN1 and AHN2 do not provide such information. Considering that different pulse densities can potentially have different effects on LiDAR metrics in different habitat types, we have performed this sensitivity analysis in five major habitat types (i.e. dunes, marshes, grasslands, shrublands, and woodlands) within Natura 2000 sites in the Netherlands. For each habitat type, we randomly selected 100 sample plots (10 m × 10 m, 500 plots in total) in which Hp95 is not NA (i.e. assuming that vegetation occurs in the plots). The detailed methodology for plot selection is provided in Appendix A. For each sample plot, the pulse density (i.e. the density of first return points) of AHN4 was systematically down-sampled to the same pulse density as AHN3, and then to 1/2 of the pulse density of AHN3 (assuming comparability with AHN2), and to 1/4 of the pulse density of AHN3 (assuming comparability with AHN1). For systematic down-sampling, we used a methodology that we recently developed (see Appendix B of Kissling et al. 2024), i.e. the first return points were first sorted according to their GPS acquisition time (from earliest to latest) and then down-sampled to the different densities. For instance, for woodlands, we have down-sampled pulse density from 25 pulses/m^2 (AHN4) to 14 pulses/m^2, 7 pulses/m^2, and 4 pulses/m^2, respectively. We then calculated the 25 LiDAR metrics for the original AHN4 point cloud and for the down-sampled point clouds. Our analysis revealed that almost all LiDAR-derived vegetation metrics in all habitats are robust to varying pulse densities at 10 m resolution, even when calculated with strongly down-sampled pulse densities of $\leqslant$ 4 pulses/m^2 (see Figure B1–B5 in Appendix B). The exception was canopy cover ("Density_above_mean_z") and Shannon index ("Entropy_z") which markedly decreased with lower pulse densities in all habitat types, and the coefficient of variation of vegetation height ("Coeff_var") in grasslands and shrublands (Appendix B). Some metrics in grasslands also showed larger variability with down-sampled pulse densities.

This sensitivity analysis provides comprehensive insights into the effects of pulse densities on metric calculations across different habitat types and helps to make our data ecologically interpretable. From this we derived usage notes for users to interpret the LiDAR metrics. We have included the sensitivity analysis and our interpretation into the revised manuscript under Sect. 3.4 "Limitations and usage notes". See Sect. 3.4.4 "Sensitivity analysis" and Appendix B for more details. We have also added a few sentences in the discussion to address this point. See lines 703–712.

2. Leaf phenology: If possible, the authors should provide a timestamp/acquisition time for every 10 m x 10 m pixel. There are vector files with information on flight lines available for AHN2-5 online, so maybe these could be rasterized? Some of the layers are likely incomplete, but having information on flight time for the pixels down to the month would make the dataset very valuable.

Response: We agree that providing a timestamp raster layer at 10 m resolution would be beneficial for comparing the time of acquisition for each grid cell among the datasets and generated properties. We have now processed the flight line information to timestamp raster layers for both AHN3 and AHN4 at 10 m resolution across the whole Netherlands. For both AHN3 and AHN4 surveys, a flightline vector layer is available with complete flight year/month/date information across the country. Note that the flightline layer of AHN2 is not complete and only a small portion of available flightlines has information on flight year and month. Hence, we eventually did not include the AHN2 flightlines in the processing. For the ongoing AHN5 survey, the flight timestamp raster layer can be included in the future. For AHN3 and AHN4, we first downloaded the flightline vector layers from https://www.ahn.nl/dataroom, and then generated a buffer zone around the flightlines using the function “Buffer” in ArcGIS Pro with the setting of a distance (on both sides of each flightline) of 300 m for AHN3 and 700 m for AHN4, and dissolved the neighboring buffer zones if they had the same flight time. The distance of the buffer zone was set based on the distance between two flightlines for the target AHN survey. We then rasterized the generated buffer zone polygons into raster layers at 10 m resolution. For areas with multiple flightlines overlapping, we assigned the latest flight date to the raster pixel to be consistent with the flight year maps provided by AHN (see Fig. 1). We suggest that users take the surrounding pixel information into account when investigating overlapping areas. We make the generated timestamp layers for AHN3 and AHN4 available in the same data repository as the data products.

We have added related information in section 3.3 Auxiliary data, see lines 354–369.

3. Quantitative guidelines: Finally, the paper should provide clear **quantitative guidelines** for researchers or practitioners on what kind of differences are interpretable ecologically. E.g., if I notice a change in height of 1m between 2017 and 2022, is this a real height change, or does this fall within the uncertainty due to laser instrumentation/DTM derivation? Vegetation growth can be slow (0.1-0.5 m/ year), so it is important to be able to separate noise/artefacts from actual change. Cf. also the second main comment.

Response: We agree that small height changes (e.g. within 0.5–1 m) can be difficult to distinguish from noises or uncertainties introduced by systematic errors or DTM derivation. Given the vertical accuracy of AHN2–AHN4 (i.e. 5–15 cm), classification related errors, and the potential influence of acquisition time of the datasets, we suggest that small vegetation changes (e.g. less than 0.5–1 m) should be interpreted with caution. Such small height changes can be influenced by vertical height uncertainties, low vegetation points being wrongly classified as ground points, or differences in leaf phenology due to varying data acquisition times rather than representing real vegetation changes. When comparing vegetation changes between the AHN3 and AHN4 metrics, users can make use of the flight time raster layers to take vegetation phenology differences into account. Based on our sensitivity analysis, we also suggest that users should be aware that some LiDAR metrics from open and heterogeneous habitats such as grasslands and shrublands might be less robust to varying point and pulse densities than those from dunes, marshes and woodlands.

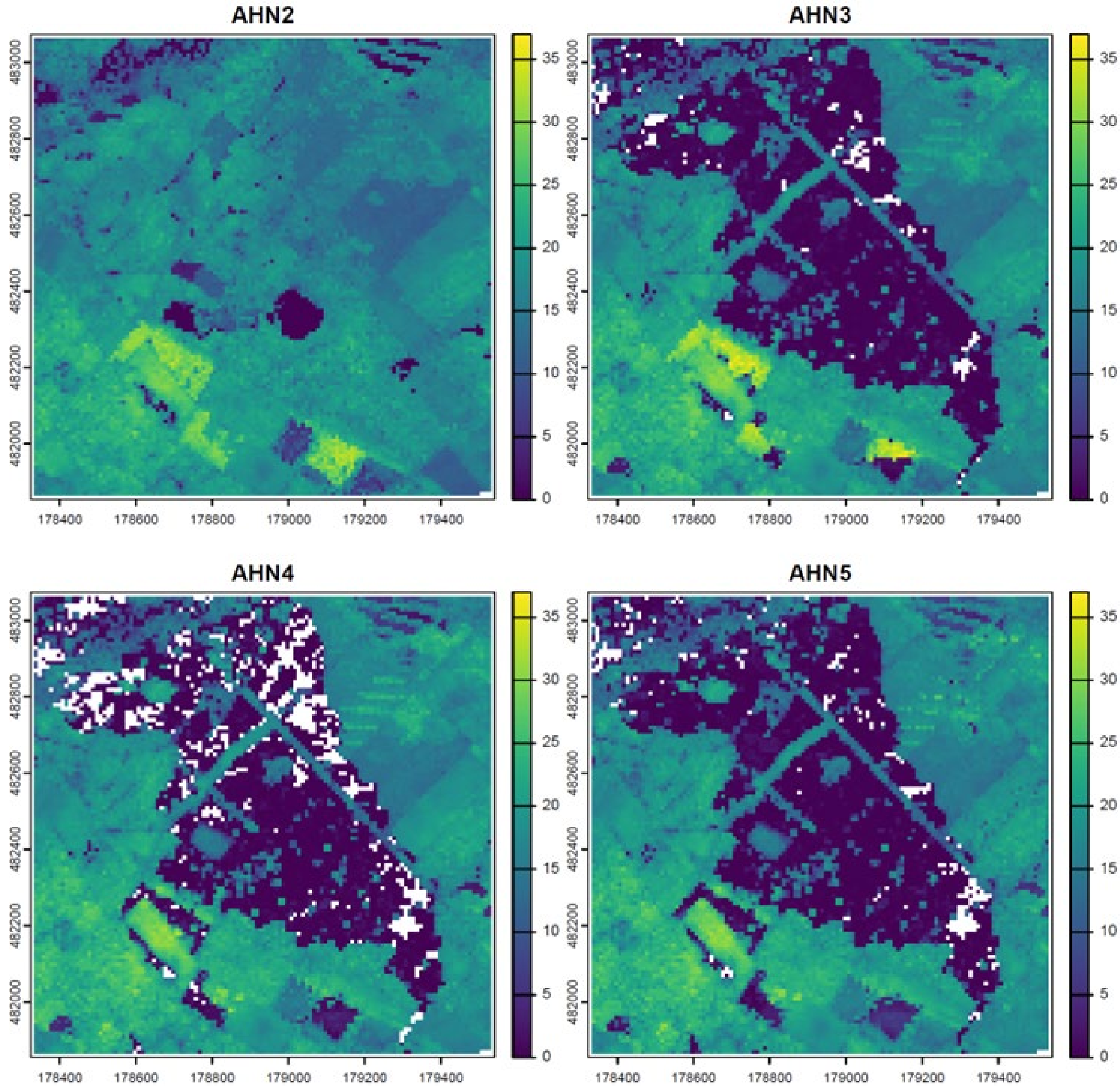
We have added this information in the revised manuscript under Sect. 3.4 “Limitations and usage notes”. See lines 525–534.

Main comment 2: Independent comparison and NA values

I carried out a quick comparison with our own laser scanning pipeline, which we have previously tested for robustness (Fischer et al., 2024, Methods in Ecology and Evolution (https://doi.org/10.1111/2041-210X.14416). I will call this the LAStools pipeline, and the author’s pipeline the Laserfarm pipeline. I have uploaded the products of this comparison to Zenodo so that the authors can compare it to their results: https://zenodo.org/records/14722001.

I only carried out a simple comparison: I compared the 95th percentile of canopy height at 10 m resolution from two CHMs produced via the LAStools pipeline (“chm_lspikefree.tif” and “chm_tin.tif”) with the Laserfarm hp95 product at the site in case study 1. The main findings are:

1. NA values: The products in the Laserfarm hp95 sometimes seem to have a considerable amount of NA values. Unfortunately, these NA values are not consistent across AHN surveys, so comparisons of canopy height change may vary depending on how these NA values are dealt with across surveys: when I ignored these differences and simply calculated the difference in canopy height means at the study site, I would get a height loss of -3.84 m from AHN2 to AHN3, a height increase of 0.45 m from AHN3 to AHN4, and then again a loss of -0.57 m from AHN4 to AHN5. When only considering areas that were not NA in any of the surveys, the height changes were as follows: -3.30 m, -0.03 m, -0.12 m, so differences of up to 0.5 m. The authors should either try to remove the NA values consistently from all products (this should be possible, as shown by the products I derived with the LAStools pipeline), or provide a mask and a clear guideline on how to deal with them. Cf. the attached pdf for a visualization of the NA values.



Response: Thank you very much for carrying out such a comparative analysis. The NA values result from metric calculations in grid cells where no vegetation points are available. Those areas are often bare ground, buildings or water bodies, which should be excluded from vegetation structure assessments. We have now generated a NA value mask for each AHN dataset, which can be used for masking out areas with potentially no vegetation. We have added detailed information into the manuscript (see lines 485–491). We did not choose to remove the NA values consistently from all products as not all users might look into temporal changes across the AHN datasets. Instead, we provide individual NA value masks for each AHN dataset. This benefits also users who only want to use one scene of raster layers for their analysis, e.g. if they want to remove the non-vegetation areas from that specific period.

Users can also combine the NA value masks when comparing vegetation changes across multiple AHN datasets. We also noted in the revised manuscript that some of the NA values can result from misclassifying very low vegetation into ground points. Hence, "no-vegetation areas" as derived from the NA value masks may differ from real land cover. See lines 491–494.

2. Differences between pipelines: The products from the LAStools pipeline are CHMs, whereas hp95 is derived from the point cloud, so we expect some differences, but they should not be massive, as both are assessing top canopy height. However, I still found clear quantitative and qualitative differences. With the chm_lspikefree.tif (the closest equivalent to top height) I found a height loss of -3.94 m from AHN2 to AHN3, a further loss of -0.36 m from AHN3 to AHN4, and then a minimal gain of -0.05 m to AHN5. The pattern was similar with chm_tin.tif, but with smaller shifts: -2.51 m, -0.47 m, then a minimal loss of -0.02 m. The absolute differences between chm_tin.tif and chm_lspikefree.tif are expected (cf. Fischer et al. 2024), but both suggest a clear height loss from AHN2-AHN3, a smaller, but clear loss from AHN3-AHN4, and then stabilization between AHN4-AHN5. This is in contrast to the results described above for hp95, which shows different patterns.

Response: Thank you very much for the detailed comparison of the differences between pipelines. Differences between pipelines can result from various reasons. Evaluating the vegetation change over a rather large area (1200 m × 1200 m in the conducted comparison) solely by an average height gain/loss can be problematic. The presence/absence of NA values as mentioned above can already have a strong influence on such average height gain/loss calculations. Moreover, the layers "chm_lspikefree.tif" and "chm_tin.tif" derived from LAStools are calculated at a 1 m resolution, whereas the "Hp95.tif" derived from the Laserfarm workflow is calculated at a 10 m resolution. Directly comparing those products with different resolutions is problematic because the top height in a 10 m pixel will likely differ from the average top height of the 100 pixels with 1 m resolution within the same 10 m pixel. A meaningful comparison should thus be calculated at the same resolution.

My main takeaway from this comparison is that change analysis in forest ecosystems is tricky and that average differences < 1.0 m may be hard to interpret/verify, unless robustness is explicitly quantified or pulse density included in the analysis. It would be good if the authors could reflect on this more clearly in the paper and supplement the current paper with a robustness test as described in point 1.

Response: Thanks for the valuable insights and suggestions. Vegetation changes less than 0.5–1 m may indeed be difficult to interpret and verify. As a supplement, we have now added a comprehensive sensitivity analysis of the generated LiDAR metrics regarding changing pulse density. This provides insights into the robustness of each LiDAR metric in different habitat types and inform users on potential uncertainties of the data products. We now also provide two raster layers (at 10 m resolution) that contain flight timestamp information for the AHN3 and AHN4 datasets, so that users can also take the time of acquisition into account when comparing vegetation changes across the datasets. Moreover, we also provide individual NA value masks for each AHN dataset, assisting users to remove NA values from further analysis.

We have addressed the abovementioned questions one by one in detail (see responses above) and have added related information in the revised manuscript accordingly.

Line-by-line:

14: I appreciate that the authors calculated structural metrics also for AHN1, but it seems to me that the data from AHN1 cannot really be interpreted for ecological purposes due to their low and highly varying quality. The authors suggest as much in the text. I think a more accurate description here would be 2007-2022, and then mentioning in the text that, theoretically, AHN1 is also available, but should only be used with great caution.

Response: We agree that the AHN1 dataset has a rather poor quality, which limits its use for ecological applications, especially if one aims for a high accuracy at a fine (e.g. 10 m) resolution. We have now reformulated the text in the revised manuscript, e.g. at the end of the introduction (see lines 139–140) and in section 2.2 (see lines 163–165).

27-29: Impressive data volumes

Response: True.

39: Great!

Response: Thanks.

55: Minor comment, but I would disagree that laser scanning is more "direct" than field measurements. It involves scanners, processing point clouds or waveforms, making assumptions about their properties, then aggregating with indices, etc. Maybe another word would be appropriate: "more precise measurements"? "more robust"?

Response: We have changed the wording to "airborne laser scanning has enabled precise and spatially contiguous measurements of ecosystem structural properties…".

61: the type of the object ("Classification") is not recorded by the laser sensor. This is post-processing and involves many assumptions. I would remove this here.

Response: Removed.

99: I agree that terrain modelling is the primary aim, but I would not call a DSM a terrain model, maybe remove and only mention DTMs

Response: We now better specified this in the text, i.e. mentioning both terrain and surface elevation models such as Digital Terrain Models (DTMs) and Digital Surface Models (DSMs).

109-126: I fully agree. Particularly, for multi-temporal lidar, harmonization is key.

Response: Agree.

154-155: The Dutch campaigns are impressive, but two clear caveats need to be mentioned in the study: 1/ Winter acquisitions are not ideal for vegetation structure assessments, because they will make it harder to detect small shrubs and lead to different estimates of canopy height between deciduous/broadleaf trees and evergreen/conifer trees. 2/ December to April is actually quite a long period. I imagine that April is already springtime in the Netherlands, with many deciduous trees growing their first leaves and thus capturing more returns higher in the canopy. A repeat study with the first scan in December and the second scan in April might thus (wrongly) conclude that a forest has increased in canopy height/closure.

Response: We fully agree with the reviewer that the acquisition times of the different AHN surveys are not ideal for measuring vegetation structure change, especially for areas in which the data collections were conducted in different months (ranging from December to April). We have therefore supplemented the data products now with newly calculated flight timestamp raster layers so that users can take these differences into account (see lines 354–369 and Sect.7 Data availability).

158-159: While I agree that it’s great to provide AHN1 as well, maybe this should state that it is not suitable for vegetation assessments? Under the worst circumstances (1 point per 16 sqm), analyses would be highly biased both due to DTMs and CHMs.

Response: Agree. We have now added a sentence in the revised manuscript to make this more clear. See lines 163–165.

161: Is this point or pulse density? I think pulse density is generally preferable as it is less instrument dependent

Response: On the official AHN website (https://www.ahn.nl/kwaliteitsbeschrijving), “Puntdichtheid” is used to describe the datasets, which translates to “point density”. For consistency, we also use point density provided by the AHN throughout the manuscript. For comparison, we have now also produced the pulse density raster layers for the AHN3 and AHN4, and conducted the robustness test of the generated LiDAR metrics based on pulse density.

184: Nice table, very clear, thanks!

Response: Thanks.

189-207: This sounds like a great pipeline, but I am missing key information on robustness here. Just from reading this, it is not at all clear to me that the pipeline produces “consistent […] geospatial products from different ALS data.” A couple of points come to mind:

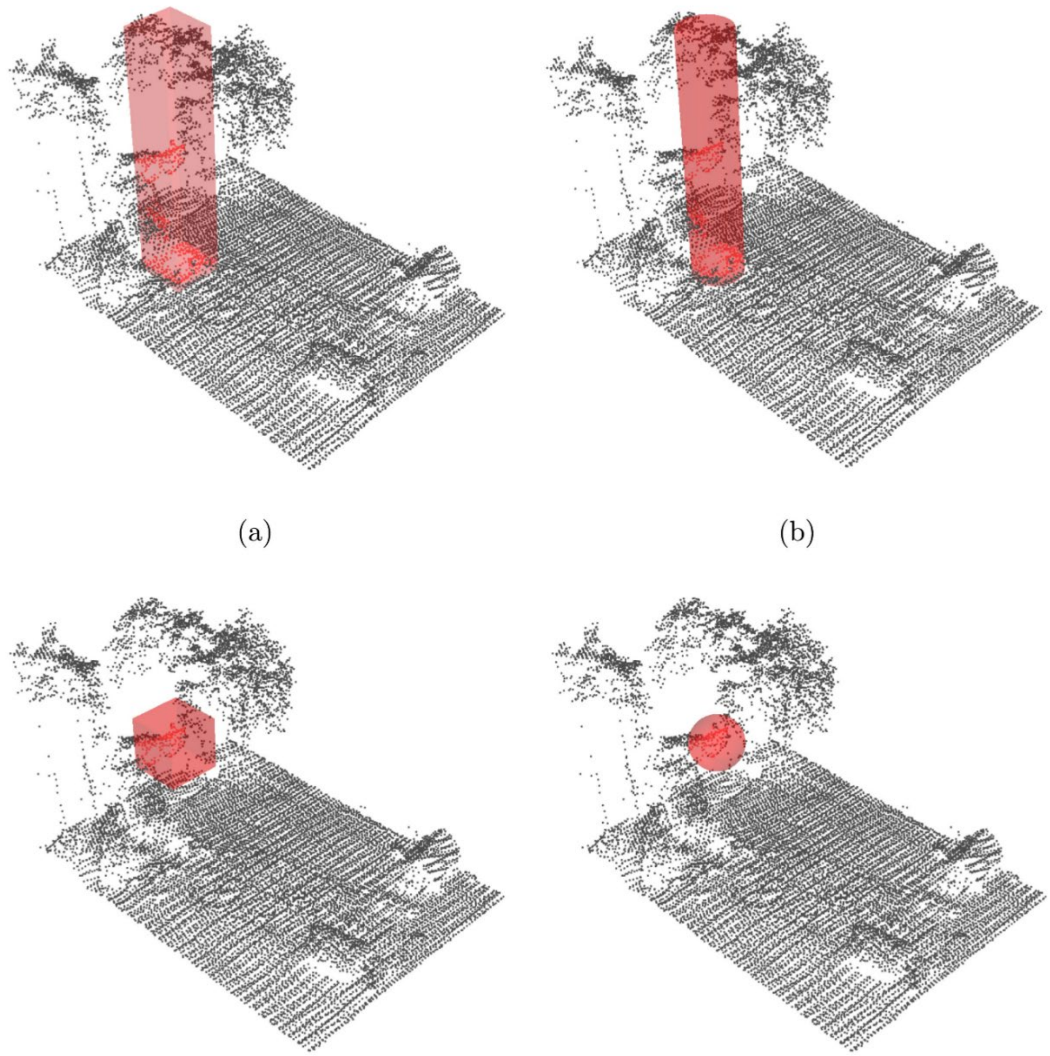
- Have you tested how the pipeline performs when laser scan pulse density (not point density) is degraded systematically, e.g., from 10 to 5 to 2 to 1?
- Maybe the description of the normalization is incomplete, but, as it is currently described, it seems non-standard and prone to large biases. The standard approach is to (robustly) ground-classify points, then interpolate a DTM, and use the inferred heights to normalize the point clouds. Using the lowest points, by contrast, seems prone to introduce a lot of noise, plus: what happens to cells that do not have a lowest point? Are cells with a single point by default classified as 0 height?
- I do not by default understand what an "infinite square cell" is, and I think this should be explained without needing to refer to another publication

Response: Thanks for the comment. We address each point one by one below:

- We have added a comprehensive robustness analysis using systematically down-sampled pulse densities in five major habitat types (dunes, marshes, grasslands, shrublands, and woodlands). These results are consistent with a previous robustness analysis of the Laserfarm workflow which was conducted by down-sampling point densities in woodland habitats only (see Appendix B of Kissling et al. (2024): https://doi.org/10.1016/j.ecolind.2024.112970). The results of our new robustness analysis are presented in Sect. 3.4.4 Sensitivity analysis and Appendix B.

- We apologize for not being very clear about the description of the normalization step. In the Laserfarm workflow, a Digital Terrain Model (DTM) is constructed using the lowest point within a given grid cell (in our case, 1 m × 1 m grid cell is used). The lowest point in each cell is taken as the height of the DTM. Each point in the cell is then assigned a normalized height with respect to the derived DTM height (Meijer et al., 2020). It results in strictly positive heights and smooths variations in elevation on scales larger than the cell size. When there are no points within a cell, the cell will be assigned as NA. When there is only one point in a cell (1 m × 1 m), it will be treated as the lowest point and assign normalized height as 0. Since the final feature extraction was based on 10 m × 10 m grid cells, any normalization biases caused by single point within a 1 m × 1 m grid cell is negligible. Nevertheless, we agree that mentioning more details of the normalization method and potential biases is useful, so we made the text more explicit in the revised manuscript (see lines 475–479 and lines 738–740). We have also made the description of the normalization step more complete (see lines 200–204).

- In the Laserchicken software (i.e. the software on which the Laserfarm workflow is based on), four options are provided for defining volume geometries, i.e. for defining the neighboring points for feature calculation (see Fig. 2 below). Fig. 2a illustrates an "infinite square cell", which is a 3D square column with a base area of 10 m × 10 m and an infinite Z value. We have now made the description of the "infinite square cell" more clear in the revised manuscript (see lines 206–207).



(a) (b)

(c) (d)

Fig. 2. Examples of volume geometries (red) available to define neighborhoods (shown as red points; enclosed): (a) an infinite square cell, (b) an infinite cylinder, (c) a cube, and (c) a sphere (Meijer et al., 2020).

214: So you use ground classification for normalization? I am confused now. Cf. my points above regarding normalization

Response: For AHN1 and AHN2, the “gefilterd” (ground) and “uitgefilterd” (non-ground) were retiled separately and then merged together (under the same tile number) and used for normalization and feature extraction. However, the classification values were not used for normalization, but only for filtering points during feature extraction. We have corrected this in the revised manuscript, see line 211.

210-212: In terms of robustness, relying only on provided ground classifications puts a lot of trust into pre-existing classifications and their comparability across campaigns. They may be good, but can you guarantee that the same algorithms were used for ground classification in AHN2 and AHN4? Have you assessed this.

Response: We could not find any documents or information about the algorithms that were used for ground classification for AHN datasets. It is only mentioned in the official AHN website that AHN datasets were first classified automatically and then followed by manual corrections for noises and errors. This was done by commercial companies and no detailed quality report is publicly available, and it is therefore difficult to assess the accuracy of the pre-classification of the AHN datasets. However, a preliminary assessment of the terrain filtering process in the Dutch coastal dunes did not reveal a strong impact of the ground point pre-classification of AHN datasets on vegetation change detection (Appendix C). We have now mentioned this point in the revised manuscript. See lines 223–228.

231-250: This is a very technical description, but I think it is great, because it can serve as a guideline for other efforts to produce streamlined products like this.

Response: Thanks!

262: Why does the data volume increase under normalization? Usually it decreases, no? Do you store the data at different precision?

Response: In the Laserfarm workflow, the normalization step calculates the normalized height value and stores it (32-bit floating-point precision) as a new attribute for each individual point. Therefore the volume increases after normalization.

280-284: These are descriptions of normalization, etc., that have already been provided above. I would remove them, and move the outlier filter further upwards to the "Processing workflow" section.

Response: Done.

275: The metrics are well-chosen in terms of ecological relevance and very nicely explained in the Table. This is very nice and well thought-through. However, I don't see any reference to robustness. As the introduction of the article rightly points out, harmonizing data across different laser sensors and campaigns is a major challenge. To ensure robustness, structural metrics should also be selected by how robust they are with regard to pulse density. Metrics that I would suspect of being particularly sensitive to laser instrumentation are PPR, the Shannon index, and any of the densities of vegetation points below 3 m (below 1, 1-2, 2-3). Minor inaccuracies in ground classification could introduce huge biases/errors. Further candidates for sensitive metrics would be Kurtosis, Roughness and the 25th percentile.

Response: Please see our response to main comment 1 about our newly implemented sensitivity analysis.

301: I like how thoroughly the paper catalogues all metrics and procedures. The authors deserve a commendation for their attention to detail!

Response: Thanks!

323-331: This is also great work. Only two comments:

- Could you use this to provide an assessment of how metrics change with pulse/point density?
- I don't know how much work this would be, but I would generally prefer a pulse density raster (i.e. first or last return density), or ideally, both pulse and point density rasters. Pulse density gives a more direct impression of sampling effort and does not confound it with the increasing power of modern laser scanners and sensors (more returns per pulse).

Response: We have conducted a sensitivity analysis of the robustness of the LiDAR metrics against changing pulse density. See our responses above and Sect.3.4.4 Sensitivity analysis. We have also generated two pulse density raster layers (at 10 m resolution) for the AHN3 and AHN4, and we additionally provide the point density layers for each AHN dataset. All are made available in the data repository together with other auxiliary data. See lines 343–348.

357: A key bit that is missing in validation is how the metrics respond to variation in pulse density. For pulse density, we would usually expect the opposite biases (i.e., much lower errors at the top, but much more near the ground, cf. Fischer et al. 2024, MEE).

Response: Please see the response to main comment 1 above. When assessing the robustness of the metrics against changing pulse density, our results showed indeed effects of lower pulse densities on the Shannon index and on the number of returns above mean height within a cell ("Density_above_mean_z"). In addition, a few LiDAR metrics (e.g. "Coeff_var") showed larger variability with down-sampled pulse densities in grasslands and shrublands, but not in other ecosystems (i.e. woodlands, dunes, marshes). Hence, there are not necessarily more biases in the metrics describing the lower strata (e.g. Hp25, BR_1_2) than those describing top canopy layers (e.g. Hp95, BR_5_20). We have added the details into the revised manuscript, see lines 518–524.

406-408: Cf. my above concerns about PPR.

Response: Our sensitivity analysis did not show strong bias in the pulse penetration ratio towards changing pulse density (Appendix B). However, we did find errors in the PPR layers where pixels that cover water surfaces were given the value of 0, which falsely indicates dense vegetation cover. This was because there were no ground points in the water surface and it resulted in 0 values during the calculation of the PPR (ratio of number of ground points to total number of points). We have therefore masked out water surfaces from PPR layers and updated the data products. This was done by masking out water areas (from TOP10NL) from the pulse penetration ratio layers using the "Extract by Mask" function in ArcGIS Pro. See lines 418–423.

467: This is a nice figure and very intuitive, but I am missing a bit the quantitative assessment and how changes in growth compare to errors.

Response: We have added some quantitative assessments of the vegetation changes in the study area. See lines 544–547.

520: Also a nice analysis and a good use case.

Response: Thanks!

513-518: Or this could be a methodological artefact. DTM models are not always super robust and small shifts by 10-50 cm might already introduce lots of noise into these assessments.

Response: Agree. We have added a few sentences in the revised manuscript. See lines 525–534.

553-555: I am a big fan of the CV, but in my experience, it will also be negatively correlated with mean height (which is intuitive, since it is computed with a division by mean height). You could also consider standardizing it between 0 and 1, as described in Lobry et al. 2023, MEE (https://doi.org/10.1111/2041-210X.14197).

Response: The standard deviation is often correlated with the mean, while the coefficient of variation (CV) – also known as relative standard deviation – is not. We have done a PCA test on the 25 metrics in a previous study using the 25 LiDAR metrics from AHN3 across the whole Netherlands (https://doi.org/10.1111/ddi.13760) and it revealed that the CV is not strongly correlated with the mean. See Fig. 3 below:

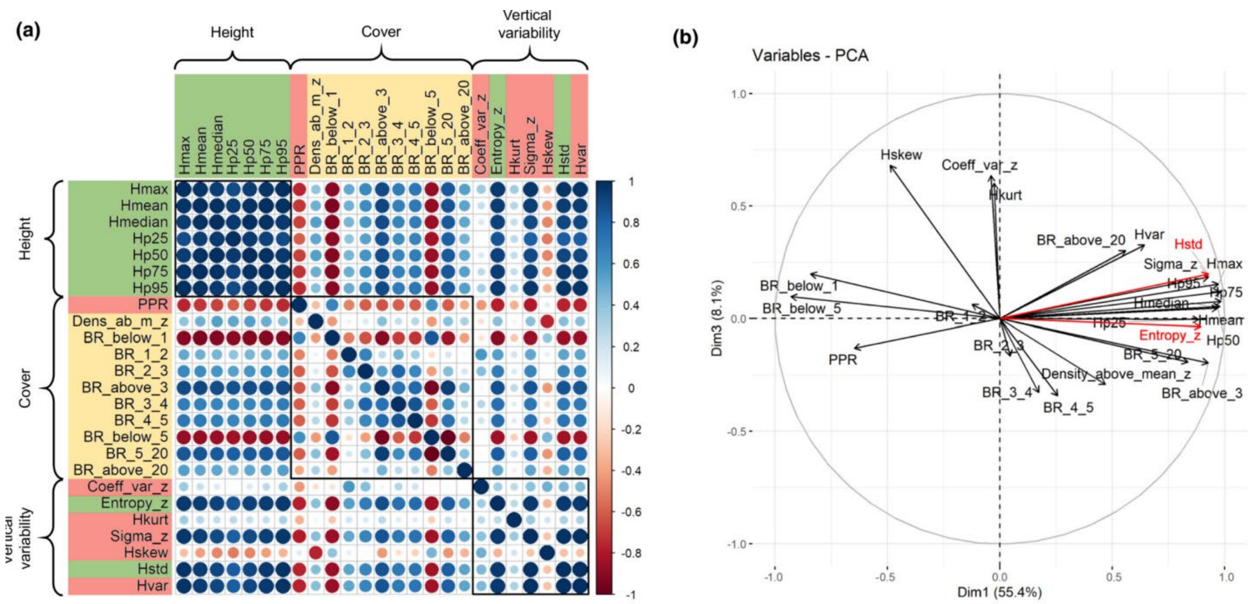


Fig. 3. Covariation among 25 metrics of vegetation structure derived from a country-wide, 10 m resolution airborne laser scanning dataset across the whole Netherlands. (a) Correlation matrix (Spearman's Rank correlation coefficients r) of metrics grouped into vegetation height, cover and vertical variability. (b) Axes from a Principal Component Analysis (PCA) explaining in total ~ 75% of variation among the 25 metrics (Kissling and Shi 2023).

573-575: I think this is a bit of a shame, because most research in ecology would likely require the DTMs with the canopy metrics, and at the same resolution. Maybe you could consider providing a few DTM metrics in the future to complement the canopy metrics?

Response: We have now generated the DTM layers for each AHN dataset at the same resolution as the provided LiDAR metrics (10 m resolution). We used the DTM layers provided by AHN and resampled them to 10 m resolution using an unweighted average method. We did the same for the DSM layers available from AHN. The DTM and DSM layers can be found as auxiliary data in the repository with provided data products. See lines 370–376.

598-603: I disagree with this. Our own research found clear problems in canopy robustness down to 2 pulses per m2.

Response: We have removed this part from our manuscript, and added the results from our sensitivity analysis. See lines 518–524 and lines 703–712.

617: This is broadly correct, but there is also the lasR package from the lidR developers that intends to be used for large-scale processing

Response: We have changed the wording in the revised manuscript to "…, lack horizontal scalability and are not specifically designed for processing large ALS datasets on cloud infrastructures with reproducible end-to-end workflows, …". See lines 723–725.

626-630: The Dutch terrain is certainly a very specific configuration and this point should be highlighted more, as more complex terrain poses many challenges. Ground classification in mountainous settings poses huge challenges, for example. This is not only where the terrain is difficult, but also where a lot of forest area remains.

Response: Agree. We have highlighted this point in the revised manuscript. See lines 475–479, and lines 738–740.

Reviewer 2

This manuscript describing a series of rasterized ALS structure products for Holland was well designed and written. It still suffers from a few areas of unclarity and/or inaccuracy which I outline below. I would also offer a few suggestions, such as

(1) I suggest "lidar" as the consensus spalling and most accepted modern usage (see: https://lidarmag.com/wp-content/uploads/PDF/LiDARNewsMagazine_DeeringStoker-CasingOfLiDAR_Vol4No6.pdf);

Response: Thank you for your suggestion. Since we have been using “LiDAR” throughout our research and publications as well as education from the beginning, and many others do so as well, we would like to keep it as it is in this manuscript for consistency.

(2) Please provide information on point cloud geolocation precision AND vertical precision.

Response: We have now added information on horizontal accuracy in addition to vertical accuracy of AHN datasets in Table 1.

(3) Please provide information on point cloud classification methods.

Response: We could not find any documents or information about the algorithms that were used in the classification of the AHN datasets. It is only mentioned in the official AHN website that AHN datasets were first classified automatically and then followed by manual corrections for noises and errors. This is typical for many ALS datasets (highlighted as one of the challenges for monitoring habitat condition from ALS in Kissling et al. 2024, https://doi.org/10.1016/j.ecolind.2024.112970) as it was done by commercial companies that do not publish detailed quality reports. Differences in pre-classification methods can potentially lead to some biases in vegetation change detection. However, a preliminary assessment of the terrain filtering process in the Dutch coastal dunes that we have now performed did not reveal a strong impact of the ground point pre-classification of AHN datasets on vegetation change detection (see our new Appendix C). Nevertheless, we suggest that this could be more comprehensively assessed in future studies, including different sites and different habitats. See lines 223–228.

(4) Some of the references (eg Asner) are a bit out of date and do not engage with theoretical developments in the literature like:

a. Coops et al. 2021. Modelling lidar-derived estimates of forest attributes over space and time: A review of approaches and future trends. Remote Sensing of Environment 260, 112477. https://doi.org/10.1016/j.rse.2021.112477

b. Cloverdale et al. 2023. Unravelling the relationship between plant diversity and vegetation structural complexity: A review and theoretical framework. Journal of Ecology 111.7 (2023): 1378-1395.

and especially for this purpose those that engage with structural typologies based on ALS:

c. Atkins et al. 2023. Integrating forest structural diversity measurement into ecological research. Ecosphere. 14(9), e4633.

and

d. Hakkenberg and Goetz 2021. Climate mediates the relationship between plant biodiversity and forest structure across the contiguous United States. Global Ecology and Biogeography. 30:2245–2258. https://doi.org/10.1111/geb.13380

Response: Thank you for your suggestions. We have included the suggested literature in the revised manuscript.

Beyond these, find some row-by-row comments and questions below:

278 - Is having the median and 50th percentile not redundant? Why have both?

Response: The 50th percentile of normalized height is indeed corresponding to the median height. We keep them both for the sake of the completion in statistical terms (e.g. max, mean, median) and percentile calculations (i.e. 25th, 50th, 75th, 95th).

278 - Why not Hp98 or Hp100? Are you not biasing results by having a low max Ht (5% below top)?

Response: We do provide the Hp100, which is the Hmax, for quantifying the top height. We provide the 95th percentile of vegetation height to quantify near-top tree canopy height while reducing the effect of outliers of the single highest points.

279 - Why 10000m? Seems too large a value. Would 1000m or even 100m not also be appropriate for Holland?

Response: The 10000 m was set to filter out outliers in the raw data, which can be noises or errors during data collection. The actual vegetation would be less than 100 m, but we did not exclude points taller than 100 m for the completion of the mapping. Yet, we provided masks for removing those artifacts, and we provide the recommendation to filter out abnormal values before using the data products for further analysis, e.g. by removing grid cells with Hp95 > 50 m, Hp95 > 40 m or Hp95 > 30 m when analyzing vegetation heights. See Sect. 3.4.1 and lines 472–474.

284 - Was there no independent DEM for verification of ground elevation? I really like the use of cadastral data for masking and verification.

Response: The DEMs provided by AHN are generated from AHN point clouds. To our knowledge, there is no publicly available independent DEM data to verify the ground elevation across the Netherlands.

287 - How accurate are these 1m vertical bins when the older imagery likely lacks that precision with so sparse a point density?

Response: We agree that the data quality of AHN1 is poor. We have clarified this point in the manuscript. See lines 139–140, lines 163–165, and lines 716–717,

Table 3 - Why 0.5m for Shannon's H and not 1m? This metric would suffer even more from vertical precision issues as noted above. Also please indicate the constraints of i. In other words i=1:? The standard answer is top HT, which in this case is biased low at Hp95.

Response: The choice of 0.5 m for Shannon's H supports it's calculation in non-forest habitats that have low-stature vegetation, such as grasses, dwarf shrubs (e.g. heath), reedbeds etc. This allows to have $i > 1$ in low-stature vegetation habitats such as dunes, grasslands, marshes etc. Most applications of Shannon's H to ALS data have been conducted in forests rather than low-stature vegetation habitats. Since we also cover a lot of non-forest habitats, we prefer 0.5 m over 1 m.

Fig. 3 - What is BR? Should be defined in caption.

Response: BR indicates band ratio. For instance, "BR_4_5" indicates the vegetation density between 4–5 m, feature name: "band_ratio_4_normalized_height_5". We have added the explanation in caption. See lines 326–327.

338 - Prior to this section, could you provide some information how (what method used) for classification?

Response: We could not find any documents or information about the algorithms that were used in the classification of the AHN datasets. It is only mentioned in the official AHN website that AHN datasets were first classified automatically and then followed by manual corrections for noises and errors. This was done by commercial companies and no detailed quality reports were publicly available. We have clarified this point in the revise manuscript. See lines 224–228.

461 - This statement on Shannon's index is incorrect. H is NOT evenness. For that you could use a metric like Pielou's J. Shannon's H is a mix of evenness and richness, where richness is the number of Ht bins and is highly correlated with Hp95 because the number of height bins is a primary parameter to the equation (the range of i, currently missing from the equation in Table 3).

Response: We have reformulated the text in the revised manuscript and made clear that the Shannon's index represents the proportion of points within 0.5 m height layers. See line 561.