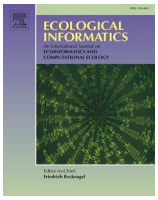
Ecological Informatics 72 (2022) 101836

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf



Laserfarm – A high-throughput workflow for generating geospatial data products of ecosystem structure from airborne laser scanning point clouds

W. Daniel Kissling [a,b,*], Yifang Shi [a,b], Zsófia Koma [a,c], Christiaan Meijer [d], Ou Ku [d], Francesco Nattino [d], Arie C. Seijmonsbergen [a], Meiert W. Grootes [d]

[a] University of Amsterdam, Institute for Biodiversity and Ecosystem Dynamics (IBED), P.O. Box 94240, 1090 GE Amsterdam, the Netherlands
[b] LifeWatch ERIC, Virtual Laboratory and Innovations Centre (VLIC), University of Amsterdam, Faculty of Science, Science Park 904, 1098 XH Amsterdam, the Netherlands
[c] Aarhus University, Department of Biology, Center for Sustainable Landscapes Under Global Change, Ny Munkegade 116, 8000 Aarhus C, Denmark
[d] Netherlands eScience Center, Science Park 402 (Matrix III), 1098 XH Amsterdam, the Netherlands

ARTICLE INFO

Keywords:
Big data
Computing architectures
Ecosystem morphological traits
Essential biodiversity variable
Macroecology
Python

ABSTRACT

Quantifying ecosystem structure is of key importance for ecology, conservation, restoration, and biodiversity monitoring because the diversity, geographic distribution and abundance of animals, plants and other organisms is tightly linked to the physical structure of vegetation and associated microclimates. Light Detection And Ranging (LiDAR) — an active remote sensing technique — can provide detailed and high resolution information on ecosystem structure because the laser pulse emitted from the sensor and its subsequent return signal from the vegetation (leaves, branches, stems) delivers three-dimensional point clouds from which metrics of vegetation structure (e.g. ecosystem height, cover, and structural complexity) can be derived. However, processing 3D LiDAR point clouds into geospatial data products of ecosystem structure remains challenging across broad spatial extents due to the large volume of national or regional point cloud datasets (typically multiple terabytes consisting of hundreds of billions of points). Here, we present a high-throughput workflow called 'Laserfarm' enabling the efficient, scalable and distributed processing of multi-terabyte LiDAR point clouds from national and regional airborne laser scanning (ALS) surveys into geospatial data products of ecosystem structure. Laserfarm is a free and open-source, end-to-end workflow which contains modular pipelines for the re-tiling, normalization, feature extraction and rasterization of point cloud information from ALS and other LiDAR surveys. The workflow is designed with horizontal scalability and can be deployed with distributed computing on different infrastructures, e.g. a cluster of virtual machines. We demonstrate the Laserfarm workflow by processing a country-wide multi-terabyte ALS dataset of the Netherlands (covering ~34,000 km^2 with ~700 billion points and ~ 16 TB uncompressed LiDAR point clouds) into 25 raster layers at 10 m resolution capturing ecosystem height, cover and structural complexity at a national extent. The Laserfarm workflow, implemented in Python and available as Jupyter Notebooks, is applicable to other LiDAR datasets and enables users to execute automated pipelines for generating consistent and reproducible geospatial data products of ecosystems structure from massive amounts of LiDAR point clouds on distributed computing infrastructures, including cloud computing environments. We provide information on workflow performance (including total CPU times, total wall-time estimates and average CPU times for single files and LiDAR metrics) and discuss how the Laserfarm workflow can be scaled to other LiDAR datasets and computing environments, including remote cloud infrastructures. The Laserfarm workflow allows a broad user community to process massive amounts of LiDAR point clouds for mapping vegetation structure, e.g. for applications in ecology, biodiversity monitoring and ecosystem restoration.

* Corresponding author at: University of Amsterdam, Institute for Biodiversity and Ecosystem Dynamics (IBED), P.O. Box 94240, 1090 GE Amsterdam, the Netherlands.
E-mail address: W.D.Kissling@uva.nl (W.D. Kissling).

https://doi.org/10.1016/j.ecoinf.2022.101836
Received 15 April 2022; Received in revised form 23 September 2022; Accepted 23 September 2022
Available online 28 September 2022

1. Introduction

Many organisms — especially animals such as birds, mammals, and insects — depend on the structural aspects of vegetation for nesting, shelter, food provisioning and foraging, and their diversity, distribution and abundance is therefore tightly linked to the horizontal and vertical heterogeneity of their habitats (Davies and Asner, 2014; MacArthur and MacArthur, 1961; Roth, 1976; Tews et al., 2004). Vegetation structure and heterogeneity also influences microclimates which has important implications for understanding climate-change impacts on biodiversity and ecosystems (Zellweger et al., 2019). Recent human-induced modifications of ecosystem structure have led to biodiversity declines, e.g. through habitat fragmentation, loss of keystone habitat structures, or through the reduction of habitat heterogeneity at the landscape scale (Benton et al., 2003; Fahrig et al., 2011; Haddad et al., 2015; Tews et al., 2004). Moreover, atmospheric nitrogen deposition, the abandonment of agricultural practices in nutrient-poor habitats, or the restoration of ecosystems through rewilding are causing changes in the structure of ecosystems (Bakker and Svenning, 2018; Fagúndez, 2012; Provoost et al., 2011). Hence, quantifying ecosystem structure in a standardized way at high spatial resolution over broad spatial extents is of key importance for ecology and conservation, e.g. in the context of ecosystem restoration and the monitoring and modelling of biodiversity (Koma et al., 2021a; Pereira et al., 2013; Valbuena et al., 2020). However, obtaining field measurements of ecosystem structure across large areas is time consuming and typically restricted to small study plots. Applications of remote sensing techniques are therefore promising because they allow to measure and monitor ecosystem structure in a spatially contiguous way and across broad spatial extents (Pettorelli et al., 2016; Skidmore et al., 2015; Valbuena et al., 2020; Vihervaara et al., 2017).

Active remote sensing techniques such as Light Detection And Ranging (LiDAR) provide a direct and accurate way to obtain detailed information on vertical and horizontal vegetation structure (Alexander et al., 2014; Bakx et al., 2019; Coops et al., 2016; Dubayah et al., 2020; Valbuena et al., 2020). For instance, LiDAR sensors installed on airplanes and helicopters are used in airborne laser scanning (ALS) surveys to capture information on canopy height, vegetation cover, vertical complexity or other 3D aspects of animal habitats (Davies and Asner, 2014; Valbuena et al., 2020; Vierling et al., 2008). Broad-scale ALS data, for instance over national or regional extents, are becoming available from an increasing number of countries across the world, e.g. Canada (Matasci et al., 2018), the US (https://usgs.entwine.io/), Australia (https://elevation.fsdf.org.au/), many parts of Europe (Table 1), and some areas in Asia and Africa (Stereńczak et al., 2020). ALS measurements use the time difference between a laser pulse emitted from an airborne LiDAR sensor and the return signal from objects on the ground (e.g. from leaves, branches and stems of vegetation, from buildings or infrastructure, or from the ground surface) to provide x,y,z coordinates and additional information (e.g. intensity, number of returns, and GPS time stamp) of these objects. To derive ecologically meaningful information, the massive 3D point clouds (typically consisting of hundreds of billions of points in a national or regional ALS survey) need to be further processed, e.g. into LiDAR metrics which statistically aggregate the 3D point cloud information within spatial units such as voxels or raster cells (Bakx et al., 2019; Davies and Asner, 2014; Meijer et al., 2020). This allows not only to map the terrain (through using LiDAR returns from ground), but also to quantify different aspects of vegetation structure (using LiDAR returns from vegetation). We follow the terminology of Valbuena et al. (2020) in the context of developing a standardized set of Essential Biodiversity Variables (EBVs) from LiDAR that could facilitate and enable large-scale biodiversity monitoring, especially in terms of variables related to ecosystem height (e.g. maximum vegetation height within a given cell), ecosystem cover (e.g. vegetation density within height layers), and ecosystem structural complexity (e.g. the vertical distribution and variability of vegetation within a grid cell).

Table 1
Examples of European open-access LiDAR point clouds derived from airborne laser scanning (ALS) surveys over a (sub)national extent. Such raw datasets (point clouds) enable the quantification of ecosystem structure at high spatial (e.g. 1–10 m) resolution by processing the multi-terabyte ALS dataset into raster layers, capturing height, cover or structural complexity of vegetation.

Country	Region	Point density	Data volume*	Download
Finland	Northern Europe	1–2 pt./m^2	4 TB	https://tiedostopalvelu.maanmittauslaitos.fi/tp/kartta?lang=en
Sweden	Northern Europe	0.25–1 pt./m^2	5 TB	https://geotorget.lantmateriet.se/bestallning/produkter/skogliglas
Norway	Northern Europe	0.2–10 pt./m^2	6 TB	https://hoydedata.no/LaserInnsyn/
Denmark	Northern Europe	0.2–25 pt./m^2	10 TB	https://datafordeler.dk/
Estonia	Northern Europe	0.2–18 pt./m^2	30 TB	https://geoportaal.maaamet.ee/eng/Spatial-Data/Elevation-data-p308.html
UK- England	Western Europe	0.5–16 pt./m^2	45 TB	https://environment.data.gov.uk/dataset/094d4ec8-4c21-4aa6-817f-b7e45843c5e0
UK- Scotland	Western Europe	1–16 pt./m^2	8 TB	https://remotesensingdata.gov.scot/data#/map
Netherlands	Western Europe	0.2–20 pt./m^2	16 TB	https://www.ahn.nl/ahn-viewer?origin=/common-nlm/viewer.html
Belgium	Western Europe	16–20 pt./m^2	25 TB	https://remotesensing.vlaanderen.be/apps/openlidar/#collapseDataDownload
Germany (partly)	Western Europe	4–10 pt./m^2	10 TB	https://www.geoportal-th.de/de-de/Downloadbereiche/Download-Offene-Geodaten-Th%C3%BCringen/Download-H%C3%B6hendaten
Switzerland	Central Europe	5–20 pt./m^2	25 TB	https://www.swisstopo.admin.ch/en/geodata/height/surface3d.html
Luxembourg	Central Europe	15–20 pt./m^2	2 TB	https://data.public.lu/fr/datasets/lidar-2019-releve-3d-du-territoire-luxembourgeois/
Spain	Southern Europe	0.5–2 pt./m^2	5 TB	http://centrodedescargas.cnig.es/CentroDescargas/catalogo.do?Serie=LIDAR
Slovenia	Eastern Europe	2–5 pt./m^2	2.5 TB	http://gis.arso.gov.si/evode/profile.aspx?id=atlas_voda_Lidar@Arso
Latvia	Eastern Europe	1.5–4 pt./m^2	10 TB	https://www.lgia.gov.lv/en/Digit%C4%81lais%20virsmas%20modelis
Slovakia	Eastern Europe	5–10 pt./m^2	8 TB	https://zbgis.skgeodesy.sk/mkzbgis/en/teren/toc?pos=48.800000,19.530000,8

* Data volume represents how much data storage is needed. It is estimated based on the number of files available in each download portal and the average size of each file.

Given the large data volumes (e.g. multiple terabytes of raw data), the processing of massive ALS point clouds is computationally demanding and often a major challenge for ecologists (Meijer et al., 2020; Roussel et al., 2020). For instance, extracting LiDAR metrics (e.g. ecosystem height, cover and structural complexity) from ALS point clouds with high resolution over national or regional extents requires performing calculations over hundreds of billions of points, posing challenges in terms of required central processing unit (CPU) time, memory capacity, data storage and data access. Moreover, the large data volumes require the re-tiling of the raw data into chunks with appropriate size, to optimize memory allocation during processing, and to take advantage of multi-core or multi-machine architectures, e.g. making feature extraction amenable to distributed computing and parallel

processing. Several software tools already exist for ALS data processing, but some of them — such as OPALS (Pfeifer et al., 2014) and LAStools (http://lastools.org/) — are not open source or contain source code and functionalities that are not publicly available (see brief review in Roussel et al. (2020)). Open source tools such as LidR (Roussel et al., 2020), FUSION (http://forsys.sefs.uw.edu/fusion/fusionlatest.html), CloudCompare (https://www.danielgm.net/cc/), and the Point Data Abstraction Library (PDAL, https://pdal.io/) exist, but typically do not provide reproducible end-to-end workflows for massive parallel scaling that can be deployed on different high performance computing platforms. A key bottleneck is the availability of free and open-source software (FOSS) tools that allow a high-throughput processing of multi-terabyte LiDAR point clouds in an efficient, scalable and distributed way. In fact, many existing software packages and tools are not capable of handling large amounts of input data, limiting their use in upscaling the LiDAR point cloud processing to broad spatial extents, e.g. for analysing fine-scale habitat requirements of threatened species (de Vries et al., 2021; Koma et al., 2021a) or EU-wide habitat condition monitoring (Pereira et al., 2022). Moreover, while some researchers with a sufficient degree of computer literacy can overcome challenges of big data management and processing of LiDAR point clouds, they typically use aggregations of custom-made scripts which have limited reproducibility. Workflows that facilitate a detailed automatic documentation linking input data to outputs while including processing parameter choices is vital for ensuring that results are reproducible. Hence, the development of modular, reproducible and scalable high-throughput FOSS workflows will increase reproducibility and enable users to handle large data volumes in a consistent and computationally efficient way.

Here, we present 'Laserfarm', a reproducible high-throughput FOSS workflow for the standardized and scalable processing of massive amounts of LiDAR point clouds from national and regional ALS surveys into raster layers of ecosystem structure (Fig. 1). The Laserfarm workflow supports interoperability and reusability (Wilkinson et al., 2016) and is designed for (1) free and open use (i.e. no restrictive license, free of charge, and with inspectable and modifiable code), (2) horizontal scalability (i.e. to execute multiple processes in parallel and to distribute the workload over multiple nodes), (3) deployment on different computing infrastructures (from single machines with multiple nodes, to clusters of virtual machines, supercomputing clusters, and cloud computing), and (4) reproducibility (i.e. automatic documentation detailing the inputs and parameters used in generating its output). We illustrate the implementation and performance of the Laserfarm workflow with a country-wide LiDAR dataset from the third Dutch national ALS flight campaign (AHN3), covering ~34,000 km^2 with a point density of ~10–20 points/m^2. The ~700 billion points and ~16 TB uncompressed data volume are processed into 25 raster layers that quantify various dimensions of ecosystem structure, including aspects of ecosystem height, ecosystem cover, and ecosystem structural complexity. Laserfarm is programmed in Python which is widely adopted in the scientific community and comes with a detailed documentation, tutorials and example implementations available as Jupyter Notebooks.

2. Materials and methods

2.1. Design principles of Laserfarm

Laserfarm (https://pypi.org/project/laserfarm/) is a reproducible, modular end-to-end workflow for efficiently extracting LiDAR metrics of ecosystems structure on distributed computing infrastructures. The four modular scriptable pipelines (Fig. 2) allow the standardized and computationally efficient re-tiling, normalization, feature extraction and rasterization of massive LiDAR point clouds into high-resolution geospatial data products of ecosystem structure. A number of design principles were central for developing the Laserfarm workflow:

First, Laserfarm is fully free and open to use. This has been achieved by basing the workflow only on existing FOSS tools such as the user-extendable, cross-platform Python tool 'Laserchicken' (Meijer et al., 2020), the Point Data Abstraction Library (PDAL), the Geospatial Data Abstraction Library (GDAL), and numerous packages hosted on the open source Python Package Index (PyPI). Moreover, Laserfarm itself is also available as FOSS, i.e. free of charge, released under a permissive license (Apache 2.0), and with source code being fully open and shared on a platform (GitHub) that promotes community engagement. Finally, Laserfarm supports standard point cloud and geo-data formats (LAS/LAZ, PLY, GeoTIFF, etc.) which makes it compatible with a wide range of other (FOSS) tools for geoscience.

Second, Laserfarm is designed for horizontal scalability. This is implemented by combining the FOSS tool Laserchicken, which provides the flexible and customizable processing of point cloud data via vectorized single process operations (Meijer et al., 2020), with the mature

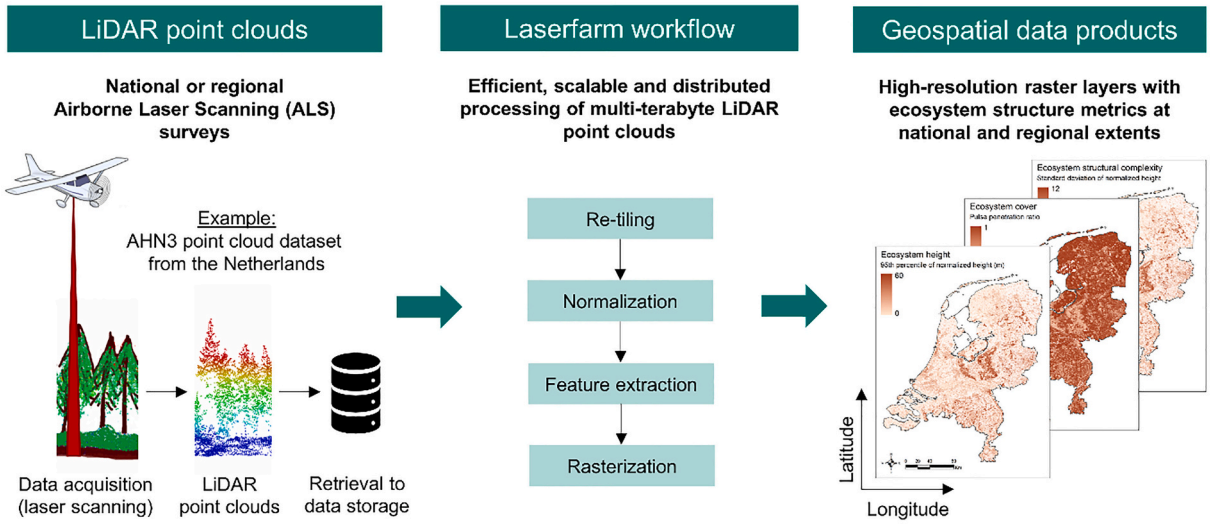


Fig. 1. The Laserfarm workflow enables the efficient, scalable and distributed processing of multi-terabyte Light Detection And Ranging (LiDAR) point clouds from national and regional airborne laser scanning (ALS) surveys into raster layers of ecosystem structure. The Laserfarm workflow is exemplified with a country-wide LiDAR point cloud dataset from the Netherlands (AHN3). Examples of specific LiDAR metrics capturing ecosystem height, ecosystem cover, and ecosystem structural complexity are provided in Table 2.

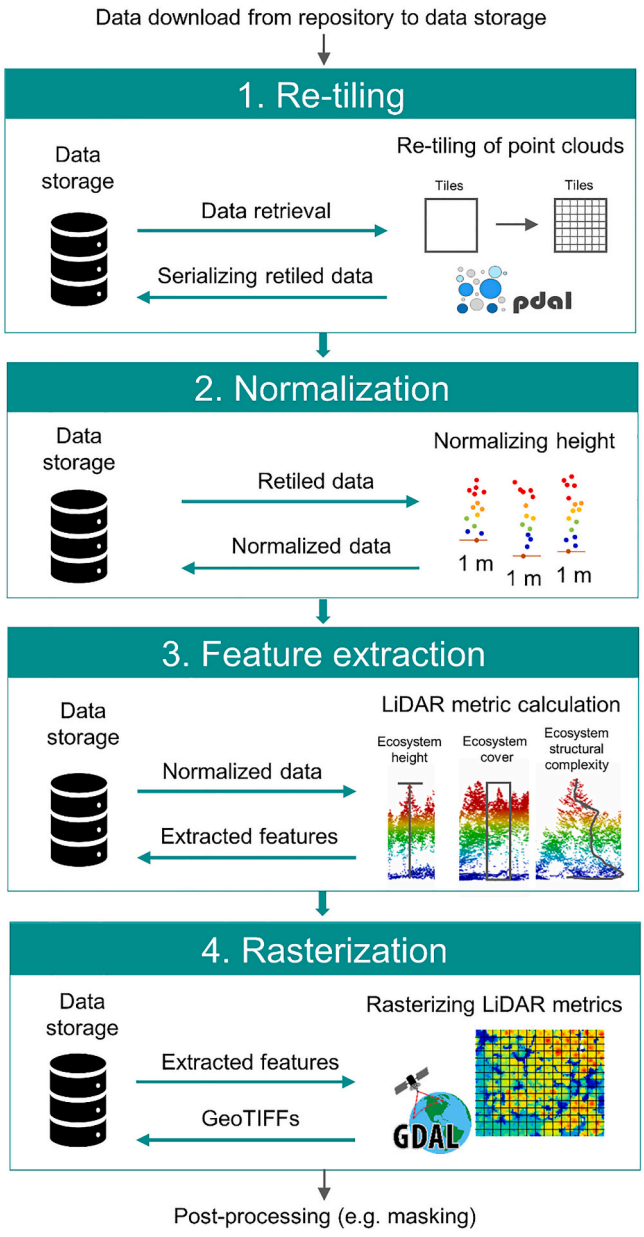


Fig. 2. Architecture of the Laserfarm workflow showing the four core modules. After downloading the LiDAR point clouds from a repository to a data storage, Laserfarm provides a standardized and reproducible framework for re-tiling, normalization, feature extraction and rasterization of multi-terabyte LiDAR point clouds into raster layers of ecosystem structure. For the normalization and feature extraction, the workflow is wrapping the 'Laserchicken' point cloud processing library whereas the retiling and rasterization requires the PDAL and GDAL library, respectively. Post-processing steps can include applying masks for water surfaces and human infrastructures (e.g. derived from cadastre or landcover information) to minimize errors related to ground detection and building heights. A detailed documentation of the workflow is available on the Laserfarm website (https://laserfarm.readthedocs.io/en/latest/) and a Jupyter notebook is available on GitHub (https://github.com/eEcoLiDAR/tutorial_ecolidar/tree/main/notebooks). All code is also hosted and freely available from GitHub (https://github.com/eEcoLiDAR/Laserfarm).

scaling abilities of the Dask library (Rocklin, 2015), which offers user friendly task/process distribution across a range of computing platforms. To avoid that input data must pass thorough a single distribution node, Laserfarm has further adopted a fully distributed approach for which only metadata describing the tasks and inputs (and the outputs) are handled by the central scheduling node. The actual processing is then fully independently distributed, allowing massive multiple read and write operations to data storage, and ensuring that input/output (I/O) scaling is not a bottleneck for the processing. As such, Laserfarm supports the highest throughput possible from storage systems to processing nodes, and is designed to process very large amounts of data efficiently, i.e. high-throughput and with high performance.

Third, Laserfarm can be seamlessly deployed on computing architectures ranging from desktop systems to distributed clusters, i.e. single machines with multiple nodes, clusters of virtual machines (VMs), supercomputing clusters, and cloud computing environments. The Dask library (Rocklin, 2015) provides adapters for different computing infrastructures and thus supports the same user-facing application programming interface (API). Laserfarm further implements data access support for both file system as well as remote storage via WebDAV (Web-based Distributed Authoring and Versioning) protocols, thus ensuring broad employability on different computing infrastructures.

Fourth, Laserfarm supports reproducibility. It automatically documents input details and choices of processing parameters. Each step of the Laserfarm workflow adds metadata entries to the dataset, detailing the inputs and parameters used in generating its output. Each output thus includes metadata documenting each step of its production.

2.2. *Workflow architecture and Laserfarm modules*

Laserfarm includes four modules (Fig. 2). The first module in the Laserfarm workflow is the re-tiling ('1. Re-tiling' in Fig. 2). The large volume of modern ALS data (e.g. country-wide LiDAR datasets with high point densities) together with efficient compression algorithms have resulted in data providers delivering their LiDAR raw data with file sizes that are often too large for general system memories on processing platforms. Hence, the original files available from LiDAR repositories need to be first re-tiled into smaller chunks for further efficient, scalable and distributed processing. After downloading the LiDAR point clouds from a repository (see examples in Table 1) to a data storage, the Laserfarm workflow splits the raw data (typically available as LAZ files) into smaller LAZ files with a user defined tile size (Fig. 2). Splitting the original LAZ files into multiple smaller LAZ files is based on the "split" functionality from the PDAL (PDAL Contributors, 2020) which allows to create smaller data objects for further processing. This takes advantage of the low-level (C++) implementation of the PDAL, and supports typical point cloud data formats such as LAS/LAZ and PLY. If the sizes of input files exceed the limit of available memory allocation, the files can be first split into smaller tile sizes before applying the re-tiling step of the Laserfarm workflow. This can be done with the laspy library from PyPi (https://laspy.readthedocs.io/en/latest/installation.html) rather than the "split" functionality from the PDAL as the latter requires a much larger memory to load the whole (original) file. Overall, the re-tiling step allows the computing infrastructure to subsequently handle the data as efficiently as possible given its available CPUs and random-access memory (RAM).

The second module of the Laserfarm workflow is the normalization ('2. Normalization' in Fig. 2). This module builds on the 'Normalize' module of the 'Laserchicken' software (Meijer et al., 2020) and normalizes the point cloud heights relative to the terrain surface. LiDAR raw data from ALS typically come with height values (z-values) that represent the absolute height rather than the height relative to the ground. The normalization therefore subtracts the ground surface and removes the influence of terrain on the height of aboveground points. Using the retiled data, Laserfarm calculates the normalized height for each individual point as the height relative to the lowest point within a grid cell size defined by the user (Fig. 2).

The third module of the Laserfarm workflow is the feature extraction ('3. Feature extraction' in Fig. 2). This module builds on the 'Features' and 'Compute Neighbors' modules of the 'Laserchicken' software (Meijer et al., 2020). >50 feature calculations are currently available in the 'Laserchicken' software (https://laserchicken.readthedocs.io/en/

latest/) of which twenty-five are particularly suited to capture aspects of ecosystem height, ecosystem cover and ecosystem structural complexity (for details see Table 2 and Section 2.5. *Data*). These LiDAR metrics can be classified into three ecosystem dimensions (height, cover and structural complexity) following a standardized framework of ecosystem morphological traits derived from LiDAR (Valbuena et al., 2020). This enables the monitoring of globally consistent variables of ecosystem structure at national or regional scales.

The fourth module of the Laserfarm workflow rasterizes the extracted features ('4. Rasterization' in Fig. 2). The rasterization includes merging and exporting the extracted features as raster layers by serializing them into data formats (e.g. GeoTIFF) that are compatible with Geographic Information System (GIS) or other software familiar to ecologists (e.g. R), based on functionality from the GDAL.

2.3. Implementation of workflow

The re-tiling module of the Laserfarm workflow needs to be configured using the functions retiling_input{} and set_grid{} (Fig. 3a). This allows users to define the specific schema for the re-tiling, i.e. the spatial extent and the number of tiles. A log file is generated for each re-tiled file which contains information of the processing steps for validation. This is done with the function validate{} (Fig. 3a) and allows users to check whether the generated (smaller) LAZ files contain the same number of points as the (larger) parent file, i.e. comparing the point count value of the original LAZ file (available in the LAZ header file) with the sum of point count values from all split files.

The normalization module of Laserfarm is implemented with the function normalization_input{} (Fig. 3b). The cell size for the normalization can be defined by the user, e.g. 1 m × 1 m ('normalize': 1, Fig. 3b) calculates the normalized z-value for all points by subtracting the height (z-value) of the lowest point within a 1 m × 1 m grid cell from each individual point in the cell.

The feature extraction module calculates the LiDAR metrics of ecosystem structure using the parameters specified in feature_extraction_input{} (Fig. 3c). The calculation of features requires to define vegetation and ground points (function apply_filter{}; Fig. 3c) and a spatial resolution (function generate_targets{} with specification of 'tile_mesh_size'; Fig. 3c), and a specific volume geometry of the point cloud (Meijer et al., 2020). Since the main focus of the Laserfarm workflow is to generate raster layers of ecosystem structure, each LiDAR metric is extracted using the point cloud around the centroids of (square) grid cells with an infinite vertical extent as the volume (Meijer et al., 2020). Note that other resolutions (for infinite square cells) or other volume geometries (infinite cylinders, cubes and spheres) can be specified because of the flexibility of the Laserchicken software to define various subsets of point cloud data (Meijer et al., 2020).

Extracting features of ecosystem structure requires to define which of the points belong to vegetation. This can be done using the pre-classification of point clouds as provided in the raw ALS datasets, typically captured using the point class standard of the American Society for Photogrammetry & Remote Sensing (ASPRS, 2019). The function apply_filter{} in the Laserfarm workflow (Fig. 3c) allows to define the vegetation points where the 'value' specifies the classification code of the ASPRS pre-classification. In our implementation, vegetation points were defined using the ASPRS classification code 1 ('Unclassified') (Fig. 3c). The feature extraction module then defines the spatial resolution (grid cell size) of the feature extraction using the function generate_targets{} and 'tile_mesh_size' (see example code in Fig. 3c). The actual LiDAR metrics are calculated by specifying the feature names in the function extract_features{} (see example 'perc_95_normalized_height' in Fig. 3c). To export the extracted metrics ('export_targets', Fig. 3c), PLY files are generated for each LiDAR metric and stored in separate folders (using 'multi_band_files': False).

The rasterization module finally exports the extracted LiDAR metrics as PLY files ('export_targets', Fig. 3c) and then stores them as single-band GeoTIFF files in separate folders (using 'multi_band_files': False, Fig. 3c).

2.4. IT infrastructure specifications

To illustrate the processing with the Laserfarm workflow, we used the IT services of SURF, the Dutch national facility for information and communication technology (https://www.surf.nl/en/ict-facilities). SURF provides access to a national IT infrastructure for the Dutch academic community, including the HPC Cloud (https://userinfo.surfsara.nl/systems/hpc-cloud) on which the computations were performed. The HPC Cloud from SURF is composed of a cluster of virtual machines (VMs) with fast CPUs and high memory nodes. We set up a cluster of 11 VMs, each VM with 2 cores, 32 GB or 64 GB RAM, and 256 GB local HDD.

Besides the HPC Cloud, we used the GRID storage infrastructure from SURF (http://doc.grid.surfsara.nl/en/latest/Pages/Advanced/grid_storage.html) as a data storage to which the raw LiDAR point clouds were downloaded from the repository of the data provider (Fig. 2). The GRID storage infrastructure was also used for managing the large amount of data (e.g. retrieving, writing and deleting files).

2.5. Data

To demonstrate the processing of LiDAR point clouds with the Laserfarm workflow, we used the ALS data from the third Dutch national flight campaign (AHN3, Actueel Hoogtebestand Nederland). AHN3 is a country-wide, open-access ALS dataset with ~700 billion points and a point density of ~10–20 points/m^2. It captures multiple returns with centimetre accuracy and has been acquired between 2014 and 2019 during the leaf-off season (between December and March). The total raw data volume is ~16 TB. The raw point cloud has been pre-processed by 'Rijkswaterstraat' (the executive agency of the Dutch Ministry of Infrastructure and Water Management) and comes with a classification code covering six classes (0: Never Classified, 1: Unclassified, 2: Ground, 6: Building, 9: Water, 26: Reserved [bridges etc.]), following the ASPRS point class standard (ASPRS, 2019). Besides the classification information, each point contains x,y,z coordinates and some additional characteristics (e.g. return number, intensity value, scan angle rank and GPS time). Other flight-related parameters such as pulse repetition rate, flight height and actual flight lines are not provided. The raw AHN3 data can be downloaded and viewed either via the Dutch geodataset platform called 'Publieke Dienstverlening Op de Kaart (PDOK)'(https://www.pdok.nl/introductie/-/article/actueel-hoogtebestand-nederland-ahn3-) or via the viewer of the 'Actueel Hoogtebestand Nederland (AHN)' (https://ahn.arcgisonline.nl/ahnviewer/). We used a custom-made script (https://github.com/eEcoLiDAR/downloadAHN) for automatically downloading the AHN3 point cloud files from the repository of the PDOK webservices (https://app.pdok.nl/ahn3-downloadpage/) to the GRID storage infrastructure.

From the AHN3 point clouds, we derived 25 LiDAR metrics that quantify various dimensions of ecosystem structure (Table 2). These metrics closely align with those that are commonly used in ecological applications and biodiversity monitoring (Bakx et al., 2019; Coops et al., 2016; Davies and Asner, 2014; Moeslund et al., 2019; Valbuena et al., 2020). We choose a spatial resolution of 10 m raster cells as this is (1) fine enough to capture the structural variability of vegetation given the available point densities of the AHN3 dataset (Table 1), and (2) of sufficiently high resolution to allow various applications in ecology and biodiversity science from landscape to regional extents (Bakx et al., 2019). The specific code and mathematical description of each feature is available from an accompanying data publication (Kissling et al., 2022).

Several of the LiDAR metrics require to define height thresholds. For instance, the density of vegetation points in defined height layers requires to define vegetation strata by setting a lower z-value (x1) and an upper z-value (x2) (see Table 2). For our example implementation, we

Table 2
Examples of Light Detection And Ranging (LiDAR) metrics capturing ecosystem structure and their implementation in the Laserfarm workflow (building on the user-extendable features from the 'Laserchicken' software: https://laserchicken.readthedocs.io/en/latest/#features). The LiDAR metrics are grouped into three key dimensions of ecosystem structure (ecosystem height, ecosystem cover and ecosystem structural complexity). Each LiDAR metric and its ecological relevance is briefly described. All metrics are calculated with the normalized point cloud. More details on metric calculation are provided in Meijer et al. (2020), on GitHub (https://github.com/eEcoLiDAR/laserchicken), and on the 'Laserchicken' documentation page (https://laserchicken.readthedocs.io/en/latest/). References provide examples of LiDAR metric use in ecological applications.

Nr	Abbreviation	LiDAR metric	Feature in Laserchicken	Description	Ecological relevance	Example references
Ecosystem height						
1	Hmax	Maximum vegetation height	max_norm_z	Maximum of normalized z	Height of canopy surface, tree tops	Bakx et al. (2019); Hyyppä et al. (2008); Lefsky et al. (2002); Maltamo et al. (2014)
2	Hmean	Mean of vegetation height	mean_norm_z	Mean of normalized z	Average height of vegetation, mean tree height	Bae et al. (2014); Höfle et al. (2012); Maltamo et al. (2014)
3	Hmedian	Median of vegetation height	median_norm_z	Median of normalized z	Vegetation height, vertical distribution of vegetation	Bakx et al. (2019); Maltamo et al. (2014)
4–7	Hp25, Hp50, Hp75, Hp95	Percentiles of vegetation height (25th, 50th, 75th and 95th)	perc_xx_normalized_height	Four metrics, capturing 25th, 50th, 75th and 95th percentiles of normalized z, respectively	Vegetation height, vertical distribution of vegetation, density in vegetation layers	Bae et al. (2014); Bakx et al. (2019); Coops et al. (2016); Maltamo et al. (2014)
Ecosystem cover						
8	PPR	Pulse penetration ratio	pulse_penetration_ratio	Ratio of number of ground points to total number of points within a cell	Openness of vegetation, canopy fractional cover, laser penetration index	Luo et al. (2015); Peduzzi et al. (2012); Yu et al. (2014)
9	Density_above_mean_z	Canopy cover	density_absolute_mean_norm_z	Number of returns above mean height within a cell	Density of upper vegetation layer	Bakx et al. (2019)
10–18	BR_below_x2, BR_x1_x2, BR_above_x2	Density of vegetation points within defined height layers (<1 m, 1–2 m, 2–3 m, >3 m, 3–4 m, 4–5 m, <5 m, 5–20 m, >20 m)	band_ratio_x1 < normalized_height < x2	Ratio of number of vegetation points in height layers to the total number of vegetation points. Height layers are defined in meter above ground (using x1 as the lower bound, and x2 as the upper bound)	Density of vegetation layers (e.g. canopy layer, understory layer, sub-canopy layer)	Bae et al. (2014); Bakx et al. (2019); (Koma et al., 2021b)
Ecosystem structural complexity						
19	Coeff_var_z	Coefficient of variation of vegetation height	coeff_var_norm_z	Coefficient of variation of normalized z within a cell	Vertical variability of vegetation distribution (ratio of standard deviation to the mean)	(Koma et al., 2021b)
20	Entropy_z	Shannon index	entropy_norm_z	The negative sum of the proportion of points within 0.5 m height layers multiplied with the logarithm of the proportion of points within 0.5 m height layers within a cell	Vertical complexity and evenness of vegetation, foliage height diversity	Bae et al. (2014); Bakx et al. (2019); (Koma et al., 2021b)
21	Hkurt	Kurtosis of vegetation height	kurto_norm_z	Kurtosis of normalized z within a cell	Vertical distribution ('tailedness') of vegetation	Bakx et al. (2019); (Koma et al., 2021b); Maltamo et al. (2014)
22	Sigma_z	Roughness of vegetation	sigma_z	Standard deviation of the residuals of a locally fitted plane within a cylinder	Small-scale roughness and variability of vegetation	Zlinszky et al. (2012)
23	Hskew	Skewness of vegetation height	skew_norm_z	Skewness of normalized z within a cell	Vertical distribution (asymmetry) of vegetation	Bae et al. (2014); Bakx et al. (2019); (Koma et al., 2021b); Maltamo et al. (2014)
24	Hstd	Standard deviation of vegetation height	std_norm_z	Standard deviation of normalized z within a cell	Vertical variability of vegetation distribution (amount of variation around mean height)	Bae et al. (2014); Höfle et al. (2012)
25	Hvar		var_norm_z			

(*continued on next page*)

Table 2 (*continued*)

Nr	Abbreviation	LiDAR metric	Feature in Laserchicken	Description	Ecological relevance	Example references
		Variance of vegetation height		Variance of normalized z within a cell	Vertical variability of vegetation distribution (dispersion around mean height)	(Koma et al., 2021b)

defined nine height layers (<1 m, 1–2 m, 2–3 m, >3 m, 3–4 m, 4–5 m, <5 m, 5–20 m, >20 m). All LiDAR metrics were calculated by using points in the class 'Unclassified' to represent vegetation points. While some non-vegetation points remain in this class, a validation with hand-labelled points showed that the misclassification rate is low and the accuracy of the derived LiDAR metrics is high (Kissling et al., 2022). For all LiDAR metrics (except the pulse penetration ratio), we excluded points from all other classes of the ASPRS pre-classification (2: Ground, 6: Building, 9: Water, 26: Reserved). Only for the pulse penetration ratio (as a measure of vegetation openness), we additionally included ground points (class 2: Ground) together with vegetation points because this metric calculates the number of ground points relative to the total number of ground and vegetation points within a grid cell (Table 2).

2.6. Statistical analysis

To illustrate the storage space requirements for applying the Laserfarm workflow to a country-wide ALS dataset, we summarized the file volumes for the (1) raw data (AHN3 LiDAR point clouds available as LAZ files), (2) re-tiled data (split LAZ files), and (3) raster layers of LiDAR metrics (GeoTIFF files). We used a Jupyter Notebook to extract the relevant information from the log files which are automatically generated by the Laserfarm workflow when running the four modules. The notebook with example code for extracting information from the log files is available from GitHub (https://github.com/eEcoLiDAR/AHN/blob/main/AHN3/10_extract_time_and_size.ipynb).

To illustrate the workflow performance, we calculated CPU times from the same log files. We quantified (1) the total CPU time of the Laserfarm workflow for generating the 25 raster layers from the AHN3 data, (2) the average CPU time for processing a single file within each of the four Laserfarm modules (re-tiling, normalization, feature extraction, and rasterization), and (3) the CPU time per file for calculating each of the 25 LiDAR metrics within the feature extraction module. For the feature extraction module, we additionally separated the CPU time of the actual LiDAR metric calculation from the CPU time needed for the pre-computations in this module, which include getting the input file from the data storage, creating the target grid, computing neighbourhoods etc. Since one of the LiDAR metrics (i.e. the pulse penetration ratio, PPR) requires ground points (in contrast to all other LiDAR metrics which only require vegetation points), we also calculated the CPU times separately for pre-computations of the PPR vs. pre-computations of other LiDAR metrics.

In addition to the CPU times, we also provide wall-time estimates, illustrating how the parallelisation and distributed processing of the Laserfarm workflow allows efficient processing of LiDAR point clouds. Since we used a cluster of 11 VMs for parallel computing, we divided the total CPU times for each Laserfarm module by 11. For two of the Laserfarm modules (normalization and feature extraction), we further run two tiles in parallel (using both cores of a VM). Hence, the CPU times for those modules were additionally divided by 2 to derive the wall-time estimates. For the other two modules (re-tiling and rasterization), the two cores of the VMs could not be used in parallel because files sizes were too large to be loaded into the memory.

Since LiDAR metrics may correlate with each other, we performed a Principal Component Analysis (PCA) to explore co-variation among all 25 LiDAR metrics. We quantified the percentage of explained variance for each PCA axis and identified the main axes of variation in ecosystem structure across all metrics. For each of the first three PCA axes, we selected the most important LiDAR metric (with the highest contribution) for illustrative purposes (i.e. country-wide mapping). We used the R package 'factoextra' and its function prcomp(), and scaled all 25 LiDAR metrics by their standard deviations. We randomly sub-sampled 10,000 (10 × 10 m) grid cells for the PCA analysis because the prcomp() function could not allocate a vector with all grid cells across all metrics.

3. Results

3.1. File volumes

For retrieving the raw data (LiDAR point clouds), we downloaded all 1367 LAZ files (each with a spatial coverage of 5 km × 6.25 km) from the PDOK webservices to the GRID storage infrastructure from SURF. Data volumes per LAZ file varied from 0.3 MB to 6 GB (Fig. 4a), with an average volume of <2 GB (Table 3). To define the re-tiling grid, we specified a square grid of 512 × 512 cells across the Netherlands with a grid cell resolution of 1 km, using the Dutch projected coordinate system (EPSG:28992 Amersfoort / RD New). Applying the retiling module resulted in 37,457 LAZ files with an average data volume of 0.15 GB per file (Fig. 4b, Table 3). After the normalization and feature extraction module had been applied, the rasterization module produced 25 country-wide GeoTIFF files with a 10 m resolution and usually <1 GB (Fig. 4c, Table 3).

3.2. Workflow performance

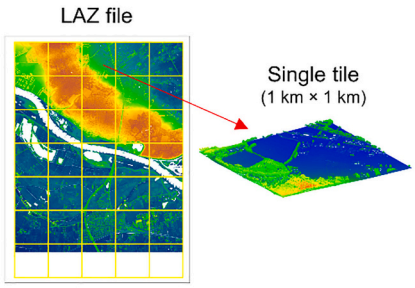
The total CPU time of the Laserfarm workflow for generating the 25 raster layers of ecosystem structure across the Netherlands at a 10 m resolution from AHN3 raw data was ~294 days (Table 4). This corresponded to a total wall-time of ~14 days given the use of 11 VMs (each with 2 cores). The feature extraction was the most time consuming Laserfarm module (Fig. 5a), followed by the normalization, re-tiling and rasterization (Table 4). Within the feature extraction module, the actual calculation of LiDAR metrics was less time consuming than the pre-computations in this module (Fig. 5a). Moreover, the pre-computations were much larger for the LiDAR metric 'pulse penetration ratio (PPR)' than for other LiDAR metrics (Fig. 5a). This is a direct result of the necessity to use both vegetation and ground points for the PPR calculation, thus making neighbourhood calculations and reading/writing of input and output files in the feature extraction module computationally expensive for the PPR.

On average, the CPU time per file was fastest for the normalization (Table 4), followed by the feature extraction and the re-tiling (Fig. 5b). Since both the normalization and the feature extraction had to be applied to 37,457 LAZ files (rather than 1367 LAZ files for the re-tiling), they resulted in a higher total CPU time (Fig. 5a). Creating the country-wide GeoTIFF file for each LiDAR metric during the rasterization took on average almost 6 h per file (Table 4), but had to be applied only 25 times (i.e. one raster layer for each LiDAR metric), hence the lowest total CPU time (Fig. 5a).

For the feature extraction, the calculation of most LiDAR metrics only took <2 s per file (Fig. 5c). The most time-consuming LiDAR metric calculation was the PPR (Fig. 5c) because it required both ground and

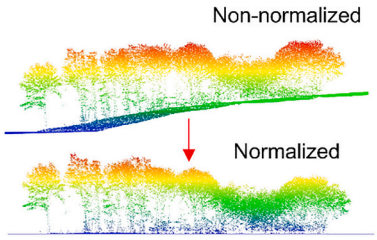
(a) Re-tiling

```
# Define the re-tiling grid
grid = {'min_x': -113107.81, 'max_x': 398892.19, 'min_y':
    214783.87, 'max_y': 726783.87, 'n_tiles_side': 512
}
# Configure the re-tiling pipeline
retiling_input = {'setup_local_fs': {
    'input_folder': remote_path_input.as_posix(),
    'output_folder': remote_path_output.as_posix()
    },
    'set_grid': grid, 'split_and_redistribute': {}, 'validate': {}
}
```



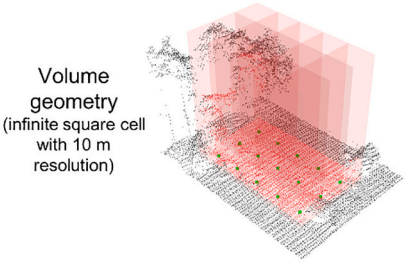
(b) Normalization

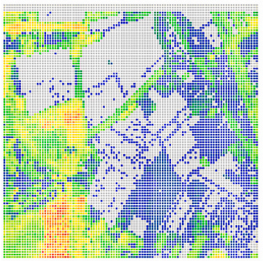
```
# Configure the normalization pipeline
normalization_input = {'setup_local_fs': {'input_folder':
path_input.as_posix(), 'output_folder': path_output.as_posix()},
    'load': {'attributes': 'all'},
    # Filter out artifically high points
    'apply_filter': {'filter_type':'select_below', 'attribute':
    'z', 'threshold': 10000.},
    # Define cell size for normalization (here 1 m × 1 m)
    'normalize': 1, 'clear_cache' : {},
}
```



(c) Feature extraction

```
# Configure the feature extraction pipeline
feature_extraction_input = {'setup_local_fs': {'input_folder':
path_input.as_posix(), 'output_folder': path_output.as_posix()},
    'load': {'attributes': ['raw_classification',
    'normalized_height']
    },
    'apply_filter': {'filter_type': 'select_equal', 'attribute':
    'raw_classification',
    # Define vegetation points based on pre-classification
    'value': 1 # AHN3 class 1 (unclassified)
    },
    # Spatial resolution for final data products (here 10 m)
    'generate_targets': {'tile_mesh_size' : 10, 'validate' :
     True, 'validate_precision': 1.e-3,
     **grid
    },
    # Define the feature name(s) and volume geometry
    'extract_features': {'feature_names':
    ['perc_95_normalized_height'], 'volume_type': 'cell',
    'volume_size': 10
    },
    # Export file
    'export_targets': {'attributes':
    ['perc_95_normalized_height'], 'multi_band_files': False,
    'overwrite': True
    },
    'clear_cache' : {},
}
```





PLY file with extracted feature (e.g. 95th percentile of normalized z for specified volume geometry)

(d) Rasterization

```
# Define output name
output_handle = 'AHN3_feat_10m_1m_veg'

# Configure the geotiff export pipeline
geotiff_export_input_nonground = {'parse_point_cloud': {},
    'data_split': {'xSub': 1, 'ySub': 1},
    'create_subregion_geotiffs': {'output_handle':output_handle},
}
```

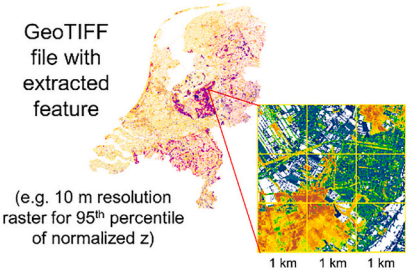


Fig. 3. Implementation with code examples from the Jupyter Notebook for each step of the Laserfarm workflow, illustrated with visualizations of the corresponding results using the Dutch AHN3 point cloud dataset. (a) *Re*-tiling defines a grid (here with 512 × 512 cells across the Netherlands using the Dutch projected coordinate system) that splits each original LAZ file into tiles of a certain size (here: 1 km × 1 km, yellow cells in LAZ file). The re-tiling pipeline also allows users to validate if the total number of points remains the same before and after re-tiling. (b) Normalization recalculates the height of each point relative to the ground surface (here relative to the lowest point within a 1 m × 1 m grid cell; 'normalize': 1). (c) Feature extraction first filters the point cloud based on a pre-classification code, then defines the spatial resolution of the final raster layers (here 10 m × 10 m), extracts LiDAR metrics based on specified feature name(s) and volume geometries, and then exports each calculated feature as a PLY file. (d) Rasterization merges all PLY files of each LiDAR metric and subsequently exports GeoTIFF raster layers (here one per metric, each covering the whole Netherlands with 10 m spatial resolution). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

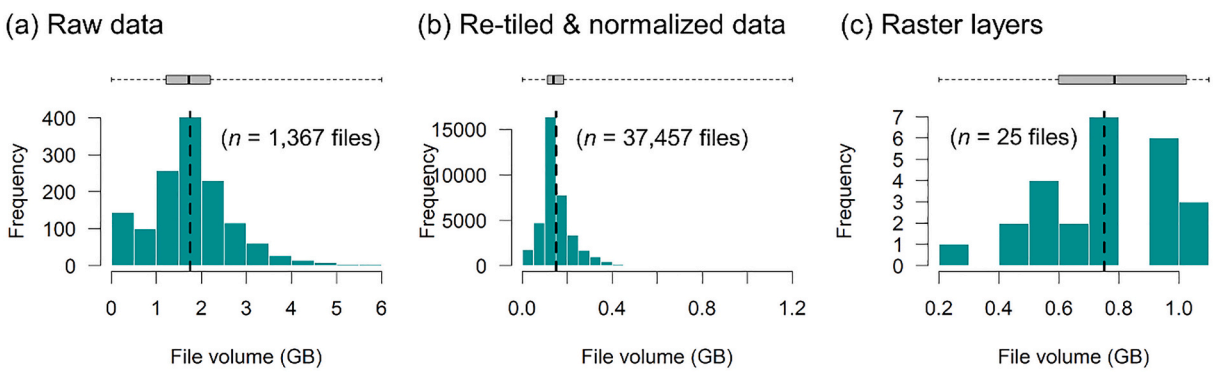


Fig. 4. File volumes of the (a) raw data from the Dutch AHN3 point cloud dataset ($n = 1367$ LAZ files), (b) re-tiled and normalized data ($n = 37{,}457$ LAZ files) after applying the re-tiling and normalization module of the Laserfarm workflow to the AHN3 raw data, and (c) raster layers ($n = 25$ GeoTIFF files) after the whole Laserfarm workflow had been applied. Boxplots show the interquartile range and median, with the whiskers extending to the data extremes. See Table 3 for summary statistics of file numbers and volumes.

Table 3
Overview of the number and volume of files used in the Laserfarm workflow as applied to the airborne laser scanning data from the third Dutch national flight campaign (AHN3). Summary statistics are provided for the (1) raw data (AHN3 LiDAR point clouds available as LAZ files from the Dutch repository), (2) re-tiled and normalized data (split LAZ files after applying the Laserfarm re-tiling and normalization module), and (3) raster layers of LiDAR metrics (final GeoTIFF files after applying all four modules of the Laserfarm workflow).

Data type	Number of files	Spatial coverage of a single file	Average volume per file (mean ± SD)	Range of volumes across files (min – max)
Raw data (AHN3)	1367	5 km × 6.25 km	1.75 ± 0.93 GB	0.27 MB – 6.00 GB
Re-tiled & normalized data	37,457	1 km × 1 km	0.15 ± 0.08 GB	0.0006 MB – 1.16 GB
Raster layers	25	Whole Netherlands	0.75 ± 0.22 GB	0.21 GB – 1.04 GB

Table 4
Performance of the Laserfarm workflow for generating geospatial data products of ecosystem structure from airborne laser scanning data of the third Dutch national flight campaign (AHN3). Total and average central processing unit (CPU) times as well as total wall-time estimates are provided. Number of files per workflow module: Re-tiling: 1367 LAZ files from AHN3; Normalization & feature extraction: 37,457 re-tiled LAZ files; Rasterization: 25 GeoTIFF files. See Table 3 for spatial coverage and volumes of files.

Workflow module	Total CPU time (days) across all files	Average CPU time (minutes) per file (mean ± SD)	Total wall-time (days)
Re-tiling	13.26	13.96 ± 9.88	1.21
Normalization	96.68	3.72 ± 1.22	4.39
Feature extraction	178.32	6.86 ± 2.44	8.11
Rasterization	5.97	343.78 ± 8.03	0.54
TOTAL	294.23	–	14.25

vegetation points. This was followed by canopy cover as the second most time-consuming metric calculation (Fig. 5c). All other LiDAR metrics took on average < 5 s per file to calculate, with ecosystem height metrics generally being calculated with <1 s per file (Fig. 5c).

3.3. LiDAR metrics

The PCA of the 25 LiDAR metrics revealed three major dimensions of ecosystem structure (Dim1–3, Fig. 6a). Together those three PCA axes explained ~75% of the variation (Appendix A Fig. A.1). PCA axis 1 (Dim1) was mainly characterized by ecosystem height (Fig. 6b), with percentiles, averages and maximum values of normalized z (i.e. Hp50, Hp75, Hp95, Hmax, Hmean, Hmedian) making the strongest contributions (Appendix A Fig. A.1). PCA axis 2 (Dim2) was mainly characterized by ecosystem cover (see Pearson coefficients of Dim2 in Fig. 6a), specifically the density of vegetation points in the lower vegetation strata (i.e. BR_1_2, BR_2_3, BR_3_4, and BR_4_5; Fig. 6b), i.e. vegetation strata up to 5 m height. Finally, PCA axis 3 (Dim3) was represented by ecosystem structural complexity (see Pearson coefficients of Dim3 in Fig. 6a), namely by skewness, kurtosis and vertical variability of normalized z value (Appendix A Fig. A.1). The three LiDAR metrics with the highest % contribution to the first three PCA axes were Hp75 (Dim1), BR_3_4 (Dim2) and Hskew (Dim3), respectively (Appendix A Fig. A.1). These three LiDAR metrics represent the three major dimensions of ecosystem structure across the Netherlands, namely the geographic variation in ecosystem height, ecosystem cover and ecosystem structural complexity (Fig. 6c–e).

4. Discussion

We present Laserfarm, a high-throughput workflow enabling the re-tiling, normalization, feature extraction and rasterization of large amounts of LiDAR point clouds into high-resolution raster layers of ecosystem structure. Laserfarm is implemented in Python, available as Jupyter Notebooks and designed with horizontal scalability for different infrastructures, i.e. the processing of multi-terabyte LiDAR point clouds through parallel execution and distribution of computations over many nodes or across a cluster of VMs. Using Laserfarm on an IT infrastructure with 11 VMs (each with 2 cores), we demonstrate the efficient, scalable and distributed processing of a country-wide LiDAR point cloud dataset from the Netherlands with ~700 billion points and ~16 TB uncompressed data volume, with an estimated wall-time of ~2 weeks. A larger number of computational nodes would further decrease the wall-time for processing a LiDAR dataset due to the high-throughput capability of the Laserfarm workflow and its seamless scalability across different computing infrastructures. It is worth noting that the actual computing time of the process might differ due to various factors, such as processing errors, VMs being offline, system errors or environment required maintenance.

The first step in the Laserfarm workflow is to re-tile the sizes and volumes of files that are available from national or institutional LiDAR repositories (Table 1). Data providers make their multi-terabyte LiDAR point clouds accessible with different file sizes because each dataset has different characteristics, e.g. in terms of data volume, point density, or spatial extent. The Dutch AHN3 dataset contained 1367 LAZ files of 5 km × 6.25 km coverage ranging in volume from 0.3 MB to 6 GB, which

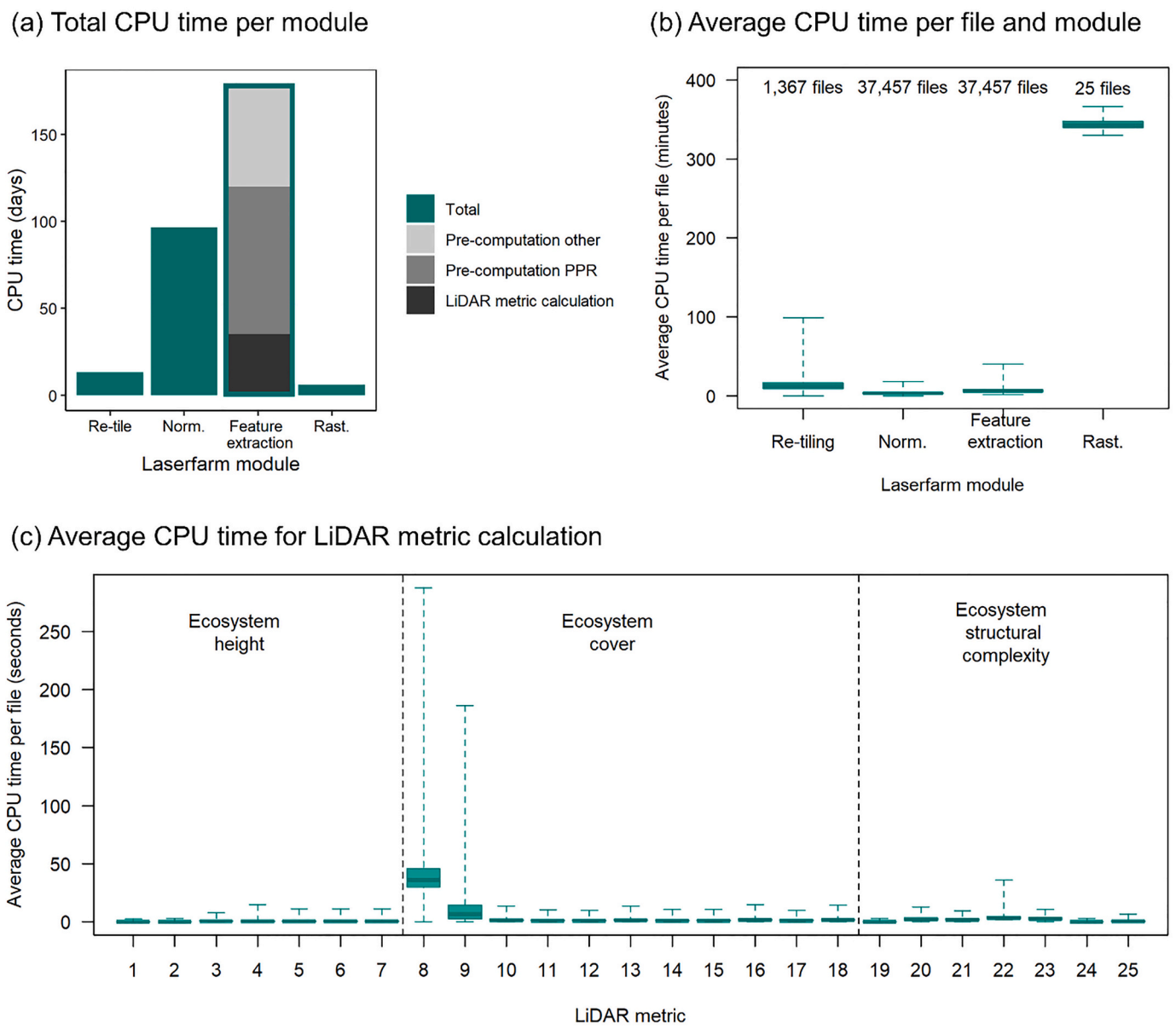


Fig. 5. Performance of the Laserfarm workflow for processing the Dutch AHN3 point cloud dataset into raster layers of ecosystem structure. (a) Total CPU time (in days) per workflow module. For the feature extraction module, the CPU time is separated for pre-computations (e.g. reading/writing files and neighbourhood calculations) and the actual LiDAR metrics calculation. Note that the CPU time for pre-computations differs between the pulse penetration ratio (PPR) and all other LiDAR metrics because the former requires both ground and vegetation points whereas the latter only vegetation points. (b) Average CPU time per file and workflow module in minutes. For the feature extraction module, the CPU times include the pre-computations and the LiDAR metrics calculation. (c) Average CPU time (in seconds) for LiDAR metric calculations within the feature extraction module (excluding the CPU time for pre-computations). The numbers on the x-axis refer to LiDAR metrics in Table 2. File numbers: Re-tiling = 1367 raw LAZ files from AHN3; Normalization = 37,457 re-tiled LAZ files; Feature extraction = 37,457 normalized & retiled LAZ files (calculating 25 metrics in total); Rasterization = 25 GeoTIFF files. Boxplots show the interquartile range and median, with the whiskers extending to the data extremes.

we re-tiled into a regular grid of 1 km × 1 km resulting in 37,457 tiles using the Laserfarm workflow. As a result of the high compression factor of the LAZ format (~10×), the large sizes of these input files (average volume of ~1.75 GB) required VMs with a large memory (64 GB RAM per node for re-tiling and rasterization, and 32 GB RAM per node for the remaining steps in the workflow). LAZ files provided by other national or institutional LiDAR repositories can differ substantially in file size and volume. For instance, the country-wide LiDAR point clouds from Spain (2nd coverage, 2015–present, 196,529 tiles, about 6 TB in total) have tiles with 2 × 2 km size and a much smaller volume of approximately 50–150 MB per file (due to the smaller coverage of each file and the lower point density compared to AHN3). The small volume of these input files may only require VMs with a small memory (e.g. 8–16 GB RAM). Users of the Laserfarm workflow can flexibly configure the computing environment, e.g. in terms of memory allocation, number of workers, and cores for each worker that are available on a specific computing infrastructure. Thoroughly testing the expected usage of memory and cores is recommended. If the sizes of input files exceed the limit of available memory allocation, the files can be first split into smaller tile sizes before applying the re-tiling step of the Laserfarm workflow. For instance, the tiles of the new (fourth) Dutch national ALS flight campaign (AHN4) are provided with the same tile extent as the AHN3 (5 km × 6.25 km size), but with a larger average volume of ~4.5 GB per file (due to the higher point density and additional attributes

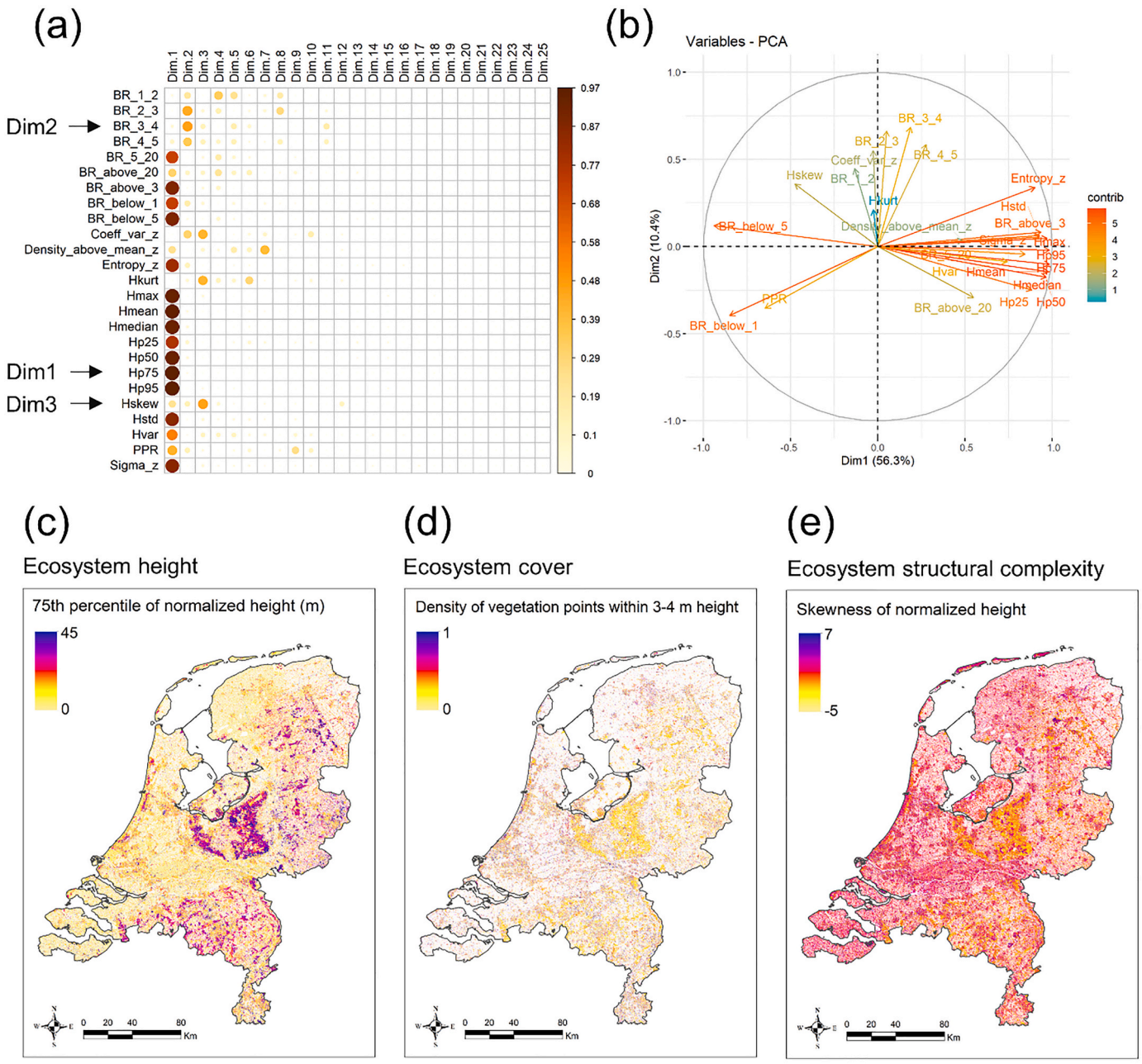


Fig. 6. Principal Component Analysis (PCA) of Light Detection And Ranging (LiDAR) metrics and their key dimensions along geographic gradients of ecosystem structure in the Netherlands. (a) Correlation plot (Pearson correlation coefficients) of all 25 LiDAR metrics (see abbreviations in Table 2). The LiDAR metric with the largest % contribution to each of the first three PCA axes (Dim1, Dim2, Dim3) is indicated with a black arrow (compare plots of contributions of variables to each dimension in Appendix Fig. A.1). (b) Co-variation and contribution of LiDAR metrics to the first two PCA axes (Dim1, Dim2) which together explain nearly 67% of variation in ecosystem structure. See Appendix Fig. A.1 for co-variation along Dim3. (c)–(e) Geographic variation of three LiDAR metrics (Hp75, BR_3_4, Hskew) that show the largest % contribution to Dim1, Dim2 and Dim3, respectively. These metrics represent aspects of ecosystem height, ecosystem cover and ecosystem structural complexity, respectively (compare Table 2).

stored in each points compared to AHN3). We provide an additional Jupyter Notebook for this which can be applied to other ALS datasets with such large tile volumes (https://github.com/eEcoLiDAR/miscellaneous/blob/splitting/jupyter_notebooks/splitting.ipynb).

The second step in the Laserfarm workflow is the normalization of each individual point which is implemented using the 'Normalize' module of the 'Laserchicken' software (Meijer et al., 2020). We used the height relative to the lowest point within a target volume, defined as a square cell with a cell size of 1 m. This is a simple and easy way to implement point cloud normalization and broadly corresponds to the most commonly used approach for normalizing non-ground points by subtracting the terrain surface from the remaining ALS returns using a derived digital terrain model (DTM) (Rapidlasso Rapidlasso GmbH, 2022; Roussel et al., 2020). Occasionally, DTM methods and ground point interpolations can lead to inaccuracies in normalized heights (Roussel et al., 2020). For instance, in small ditches or steep terrain a non-ground return (from vegetation) can be lower than the interpolated height from a DTM, resulting in negative normalized height values. Our approach of normalization is less prone to such inaccuracies because it uses the lowest height among all points in a specified cell, rather than normalizing with a DTM based on ground points only. All normalized heights are therefore positive by definition. However, a potential issue

with this normalization approach is that discontinuities in normalized heights and terrain height variability at very high resolution are less well captured in steep terrain, especially within the resolution of the defined cell size (e.g. <1 m). Independent of the normalization method, we recommend users to critically explore each dataset after normalization to detect under which circumstance inaccuracies might emerge.

A core aspect of the Laserfarm workflow is the feature extraction, which is computationally the most expensive part. We exemplify the feature extraction with a set of statistical properties of the point cloud which provide features that capture 'ecosystem morphological traits' (Valbuena et al., 2020), specifically LiDAR metrics related to ecosystem height, cover, and structural complexity. The calculation of ecosystem structure LiDAR metrics in the Laserfarm workflow makes use of the pre-classification in the raw data (LAZ or LAS files), which is typically defined using the ASPRS point class standard (ASPRS, 2019). Ground points are usually well pre-classified because a main focus of ALS surveys is terrain mapping (Assmann et al., 2022; Kraus and Pfeifer, 1998). Vegetation points are sometimes pre-classified as classes 3, 4 and 5 based on the ASPRS standard (ASPRS, 2019), representing low, medium and high vegetation, respectively. However, many ALS datasets (including the Dutch AHN3) are delivered without such a specific pre-classification for vegetation points. We therefore used the unclassified class to represent vegetation. This contains some biases (e.g. cars, poles and road fences being considered as vegetation), but the misclassification rate of vegetation points is low (~5%) and the derived LiDAR metrics show high accuracy (~90%) (Kissling et al., 2022). For the inaccuracies that remain (e.g. through ships, chimneys and cars), we provide an additional raster layer from available cadastre data in the Netherlands as a mask. However, this does not remove all inaccuracies because other human infrastructures such as railway electrification systems and powerlines are either not captured in this mask or not represented in the available AHN3 point cloud pre-classification. Users of the generated ecosystem structure data (GeoTIFF files) therefore need to explore whether errors and inaccuracies remain for their use cases or specific study sites in the Netherlands.

The final step in the Laserfarm workflow is the rasterization of the calculated LiDAR metrics (in PLY format) into raster layers (e.g. GeoTiff). These raster layers have a manageable size (e.g. ~0.75 GB per metric for 10 m resolution raster layers across the Netherlands) and can be readily used in software familiar to many ecologists (e.g. GIS or R). The spatial resolution of the raster layers is defined during the feature extraction step using the Laserchicken target volume of a square infinite cell (Meijer et al., 2020), and by defining the volume size (e.g. 10 m). Other resolutions could be flexibly defined (e.g. 1 or 5 m), but the quality, accuracy and information content of the derived LiDAR metrics can depend on the characteristics of the LiDAR point clouds (Coops et al., 2021; Gobakken and Næsset, 2008; Koma et al., 2021b; Koma et al., 2021c). The mean or upper height of vegetation derived from ALS point clouds is probably among the most robust and reliable LiDAR metrics, independent of ecosystem type (Coops et al., 2021; Koma et al., 2021c). Scalability and robustness of other LiDAR metrics may be strongly influenced by available point densities (Jakubowski et al., 2013), which vary widely among country-wide ALS datasets (Table 1). Other LiDAR data acquisition parameters such as flight height, scanner type, scan angle, and sampling design might additionally affect the comparability of LiDAR metrics (Wulder et al., 2012). Hence, accuracy assessments of LiDAR metrics from different ALS datasets (e.g. different countries or different time periods) are therefore needed within and across different types of ecosystems (e.g. forests, wetlands and grasslands). This is particularly relevant if ecosystem structure data are needed for a consistent and comparable biodiversity monitoring at high resolution over broad spatial extents, e.g. in the context of essential biodiversity variables (Valbuena et al., 2020; Vihervaara et al., 2017).

We have used the Jupyter environment in combination with the ALS data from the third Dutch national flight campaign (AHN3) together with the IT services of the Dutch national facility SURF for developing the Laserfarm workflow. Specifically, we implemented Laserfarm in Jupyter Notebooks, using a cluster of 11 VMs each with 2 cores on the HPC Cloud from SURF, and the SURF GRID storage infrastructure for managing the ~16 TB of ALS raw data, together with the intermediate LAZ and PLY files and the final raster layers. Deploying the Laserfarm workflow on other compute infrastructures (e.g. cloud computing environments such as Microsoft Azure, Amazon Web Services, or the European Open Science Cloud) will require consideration of memory allocation and performance bottlenecks during parallelization. For instance, the re-tiling of large input files requires big memory machines which are expensive and uncommon in cloud environments. This means that tiles available from ALS repositories (Table 1) might need to be split into smaller tile sizes, e.g. using the laspy library from PyPi (see methods). Applying the Laserfarm workflow to other airborne LiDAR data from different regions or countries only requires minimal modification. For instance, the coordinates of a different spatial extent have to be specified within the set_grid{} function of the re-tiling module and the classification code of vegetation points has to be specified within the feature_extraction_input{} function of the feature extraction module. LiDAR data collected on other platforms than airplanes can also be processed with the Laserfarm workflow, including 3D point clouds from terrestrial, mobile, UAV and vehicle laser scanners. The most important differences between such datasets are the point density and data volume. Hence, particular attention should be given to parameters that are closely related to point density, such as the grid cell size for the normalization and the memory allocation during the configuration.

While the Laserfarm workflow as presented here uses the Dask library (Rocklin, 2015) for efficiently scheduling the execution, each workflow module still depends on the previous one. Hence, the individual cells (corresponding to the processing steps as expressed in the structure of the Jupyter notebooks) show dependencies regarding inputs/outputs, which may result in performance bottlenecks when the Laserfarm workflow is executed on remote cloud infrastructures (e.g. each step will wait for the longest running subprocess to complete). Conceptionally, the Laserfarm workflow can also be split by input data, rather than by processing steps, with the individual data stream only being merged at the end, a model which is well suited to (commercial) cloud providers and/or more exploratory analysis at scale. Ongoing development work in this direction is focussing on containerizing the Laserfarm workflow by encapsulating it into reusable cells (e.g. as standardized RESTful API services) and testing its automated execution on remote cloud infrastructures (Wang et al., 2022). Moreover, the Jupyter Notebook of Laserfarm is currently used for developing a virtual lab in the context of a collaborative cloud virtual research environment (VRE) which provides extended functionality for composing workflows, managing the lifecycle of computational experiments, and sharing the results among a broad users community (Zhao et al., 2022). In this context, the integration of LiDAR data with other data sources (e.g. biodiversity and climate data) or analytical services (e.g. species distribution modelling pipelines) might further benefit from a federated cloud infrastructure that can address scientific computation, data analytics and heterogeneous data storage of multiple data types in one platform (Fiore et al., 2017). This could feed into the development of Digital Twins (DTs) that support advanced simulation, modelling, and prediction capabilities for monitoring and predicting environmental change and human impacts on biodiversity and ecosystems, similar to the DTs that have been proposed for climate change modelling and adaptation (Bauer et al., 2021).

5. Conclusion

LiDAR point clouds from national and regional ALS surveys are increasingly becoming available and provide opportunities for monitoring and modelling biodiversity and the structure and functioning of ecosystems. Laserfarm offers a fully free and open-source workflow for the efficient, scalable, and distributed processing of such multi-terabyte

point clouds on different computing infrastructures, including high performance computing clusters and cloud computing environments. The generated data products can be applied in ecological analyses, including the modelling of animal diversity (Davies and Asner, 2014), the prediction of species distributions and ecological niches (Bakx et al., 2019; Koma et al., 2021a), and the monitoring of a globally consistent set of ecosystem structure variables at regional scales (Valbuena et al., 2020). The extendibility and flexibility of the Laserfarm workflow also allows users to add new LiDAR metrics of ecosystems structure, or to expand the set of features to other metrics such as topographic descriptors (Assmann et al., 2022), neighbourhood calculations for detecting archaeological remnants (Inomata et al., 2020), or indicators of ecosystem condition and forest productivity (Goodbody et al., 2021). Following FOSS development best practices, we make use of an issue tracker (i.e. the GitHub Issues functionality) to collect and track user contributions such as questions, reported bugs and feature requests. We envisage that the further deployment of the Laserfarm workflow will empower the wider use and uptake of LiDAR metrics in biodiversity science, ecosystem monitoring, policy support and landscape management.

Data availability

The current version (v0.2.0) of the Laserfarm workflow is available from PyPI (https://pypi.org/project/laserfarm/) or Zenodo (doi: https://doi.org/10.5281/zenodo.3842780). A detailed documentation of the Laserfarm workflow is provided on the documentation website (https://laserfarm.readthedocs.io/en/latest/). We further provide Jupyter Notebooks (https://github.com/eEcoLiDAR/AHN/tree/main/AHN3) which allow the implementation in Jupyter (https://jupyter.org/), a web-based interactive development environment for configuring and arranging workflows in data science and scientific computing. All code of Laserfarm is also hosted and freely available from GitHub (https://github.com/eEcoLiDAR/Laserfarm). User feedback such as questions, reported bugs and feature requests are collected and tracked via GitHub (see contributing guidelines: https://github.com/eEcoLiDAR/Laserfarm/blob/master/CONTRIBUTING.md). All changes that are made to the Laserfarm workflow are also documented on GitHub (https://github.com/eEcoLiDAR/Laserfarm/blob/master/CHANGELOG.md). The Laserfarm GitHub repository further includes a tutorial structured as a Jupyter Notebook (tutorial.ipynb) which illustrates the use of the Laserfarm workflow to process a subset of the Dutch AHN3 dataset (from the retrieval of an example point cloud data file in LAZ format to the export of the extracted features to a GeoTIFF file). A second notebook (workflow.ipynb) shows the workflow employed to process the full AHN3 dataset (illustrating how the re-tiling, point cloud data-processing and GeoTIFF-exporting tasks can be configured and distributed over the nodes of a local or a remote compute cluster). Moreover, Python scripts and pipeline configuration files (https://github.com/eEcoLiDAR/Laserfarm/tree/master/examples) that have been used to test the various pipelines either on local machines or on a virtual docker-container-based cluster (https://github.com/eEcoLiDAR/dockerTestCluster) are also provided on GitHub. Versioning of the Laserfarm workflow is tracked on GitHub (https://github.com/eEcoLiDAR/Laserfarm/releases).

The raw LiDAR point clouds from AHN3 can be viewed through PDOK (https://www.pdok.nl/introductie/-/article/actueel-hoogtebestand-nederland-ahn3-) or the AHN viewer (https://ahn.arcgisonline.nl/ahnviewer/), and downloaded via the PDOK webservices (https://app.pdok.nl/ahn3-downloadpage/). A script for automatically downloading AHN point cloud files is provided on GitHub (https://github.com/eEcoLiDAR/downloadAHN). The processed raster layers of the 25 LiDAR metrics are available from Zenodo (https://zenodo.org/record/6421381).

CRediT authorship contribution statement

W. Daniel Kissling: Conceptualization, Formal analysis, Funding acquisition, Project administration, Supervision, Visualization, Writing – original draft, Writing – review & editing. **Yifang Shi:** Data curation, Formal analysis, Methodology, Validation, Visualization, Writing – review & editing. **Zsófia Koma:** Investigation, Methodology, Writing – review & editing. **Christiaan Meijer:** Software, Validation, Methodology. **Ou Ku:** Software, Validation, Methodology. **Francesco Nattino:** Software, Validation, Formal analysis, Methodology. **Arie C. Seijmonsbergen:** Funding acquisition, Writing – review & editing. **Meiert W. Grootes:** Conceptualization, Methodology, Resources, Software, Supervision, Validation, Writing – review & editing.

Declaration of Competing Interest

None.

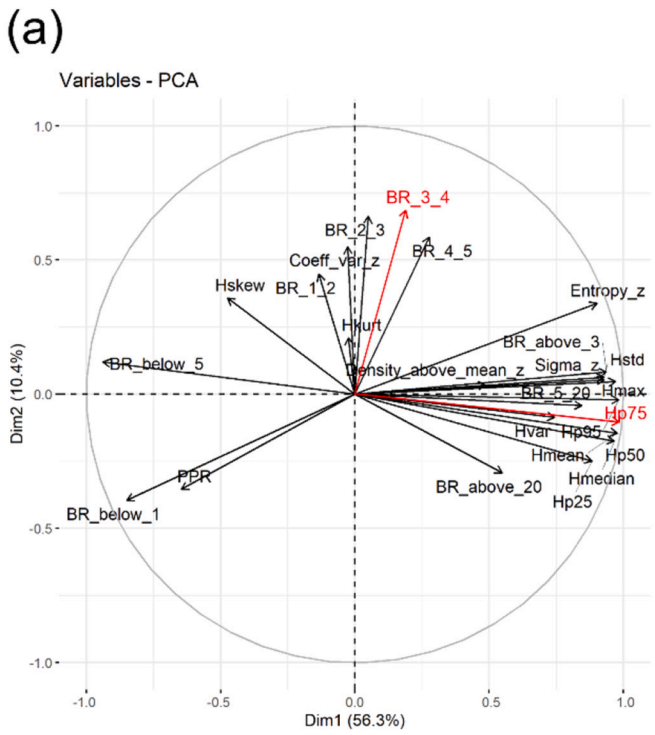
Data availability

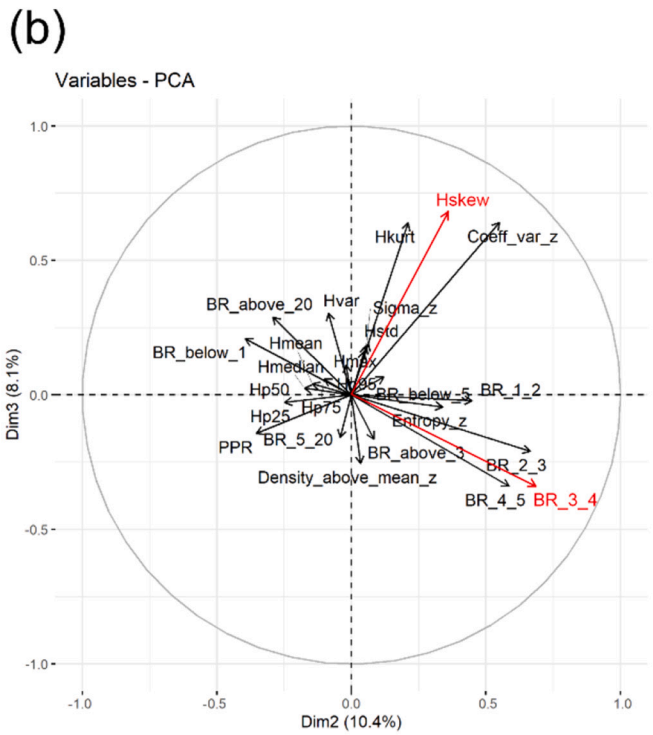
All data, code and documentation are openly accessible.

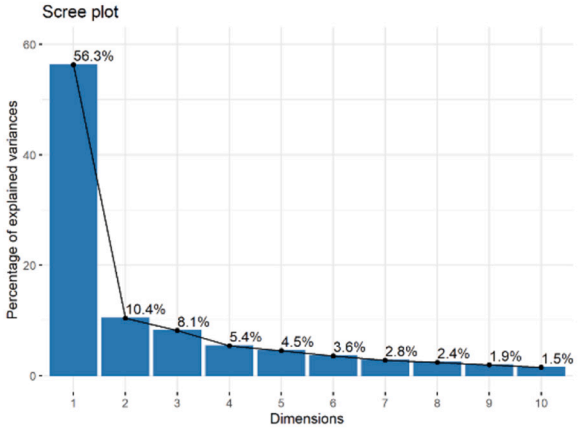
Acknowledgements

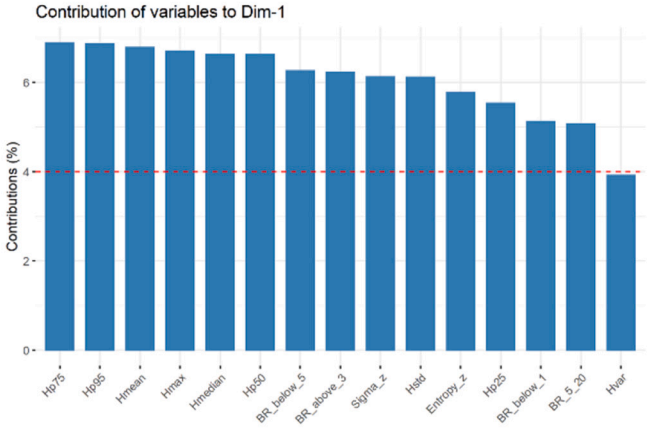
The development of the Laserfarm workflow was funded by the Netherlands eScience Center (https://www.esciencecenter.nl), grant number ASDI.2016.014, through the project ‘eScience infrastructure for Ecological applications of LiDAR point clouds’ (eEcoLiDAR) (Kissling et al., 2017). The further development of the Laserfarm workflow in the context of a virtual research environment is supported by LifeWatch (https://www.lifewatch.eu/) and a Microsoft AI for Earth Grant (AI4E-1111-Q3N7-20100806). W.D.K. also acknowledges funding from the European Union’s Horizon 2020 Research and Innovation Programme for EuropaBON (grant agreement No 101003553) which supports the development of an EU-wide framework for monitoring biodiversity and ecosystem services. The Laserfarm workflow will be deployed to support the mapping of habitat condition within the European Union (MAMBO Project, grant agreement No 101060639).

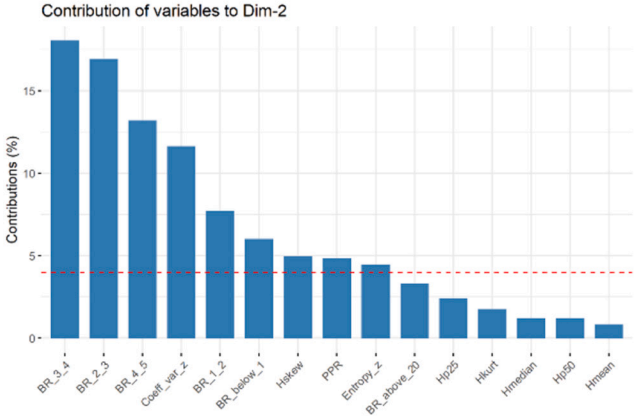
Appendix A


(a)
Variables - PCA
BR_3_4
BR_2_3
BR_4_5
Coeff_var_z
Hskew
BR_1_2
Entropy_z
Hkurt
BR_above_3
Hstd
BR_below_5
Density_above_mean_z
Sigma_z
BR_5_20
Hmax
Hp75
Hvar
Hp95
Hmean
Hp50
PPR
Hmedian
BR_above_20
Hp25
BR_below_1
Dim2 (10.4%)
Dim1 (56.3%)


(b)
Variables - PCA
Hskew
Hkurt
Coeff_var_z
BR_above_20
Hvar
Sigma_z
Hstd
BR_below_1
Hmean
Hmedian
Hmax
Hp50
Hp95
BR_below_5
BR_1_2
Hp25
Hp75
Entropy_z
PPR
BR_5_20
BR_above_3
BR_2_3
Density_above_mean_z
BR_4_5
BR_3_4
Dim3 (8.1%)
Dim2 (10.4%)


(c)
Scree plot
56.3%
10.4%
8.1%
5.4%
4.5%
3.6%
2.8%
2.4%
1.9%
1.5%
Percentage of explained variances
Dimensions


(d)
Contribution of variables to Dim-1
Contributions (%)
Hp75
Hp95
Hmean
Hmax
Hmedian
Hp50
BR_below_5
BR_above_3
Sigma_z
Hstd
Entropy_z
Hp25
BR_below_1
BR_5_20
Hvar


(e)
Contribution of variables to Dim-2
Contributions (%)
BR_3_4
BR_2_3
BR_4_5
Coeff_var_z
BR_1_2
BR_below_1
Hskew
PPR
Entropy_z
BR_above_20
Hp25
Hkurt
Hmedian
Hp50
Hmean

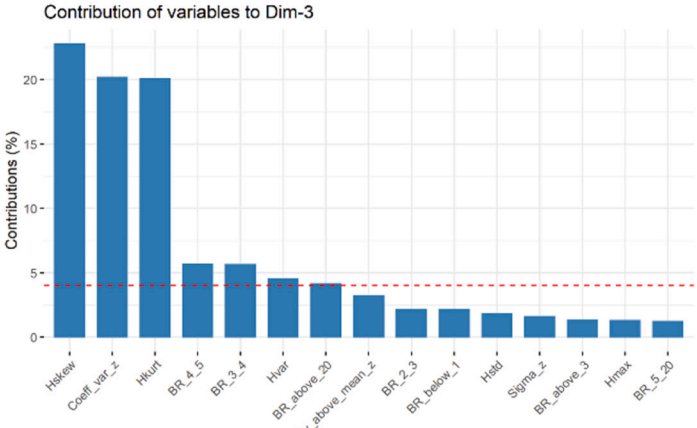

(f)
Contribution of variables to Dim-3
Contributions (%)
Hskew
Coeff_var_z
Hkurt
BR_4_5
BR_3_4
Hvar
BR_above_20
Density_above_mean_z
BR_2_3
BR_below_1
Hstd
Sigma_z
BR_above_3
Hmax
BR_5_20

Fig. A.1. Results from a Principal Component Analysis (PCA) of Light Detection And Ranging (LiDAR) metrics (see metric abbreviations in Table 2). (a)–(b) Co-variation and contribution of LiDAR metrics to PCA axis 1 and 2 (Dim1, Dim2) and PCA axis 2 and 3 (Dim2, Dim3), respectively. The metric with the highest contribution to each axis is highlighted in red. (c) Scree plot showing the % explained variance of the first ten PCA axes (dimensions 1–10). (d)–(f) Contribution (%) of each LiDAR metric to Dim1, Dim2, and Dim3, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

References

Alexander, C., Bøcher, P.K., Arge, L., Svenning, J.-C., 2014. Regional-scale mapping of tree cover, height and main phenological tree types using airborne laser scanning data. Remote Sens. Environ. 147, 156–172.

ASPRS, 2019. LAS Specification 1.4 - R15. American Society for Photogrammetry & Remote Sensing, Maryland, USA, p. 50.

Assmann, J.J., Moeslund, J.E., Treier, U.A., Normand, S., 2022. EcoDes-DK15: high-resolution ecological descriptors of vegetation and terrain derived from Denmark's national airborne laser scanning data set. Earth Syst. Sci. Data 14, 823–844.

Bae, S., Reineking, B., Ewald, M., Mueller, J., 2014. Comparison of airborne lidar, aerial photography, and field surveys to model the habitat suitability of a cryptic forest species – the hazel grouse. Int. J. Remote Sens. 35, 6469–6489.

Bakker, E.S., Svenning, J.-C., 2018. Trophic rewilding: impact on ecosystems under global change. Philosoph. Trans. Royal Soc. B: Biol. Sci. 373, 20170432.

Bakx, T.R.M., Koma, Z., Seijmonsbergen, A.C., Kissling, W.D., 2019. Use and categorization of light detection and ranging vegetation metrics in avian diversity and species distribution research. Divers. Distrib. 25, 1045–1059.

Bauer, P., Stevens, B., Hazeleger, W., 2021. A digital twin of earth for the green transition. Nat. Clim. Chang. 11, 80–83.

Benton, T.G., Vickery, J.A., Wilson, J.D., 2003. Farmland biodiversity: is habitat heterogeneity the key? Trends Ecol. Evol. 18, 182–188.

Coops, N.C., Tompaski, P., Nijland, W., Rickbeil, G.J.M., Nielsen, S.E., Bater, C.W., Stadt, J.J., 2016. A forest structure habitat index based on airborne laser scanning data. Ecol. Indic. 67, 346–357.

Coops, N.C., Tompalski, P., Goodbody, T.R.H., Queinnec, M., Luther, J.E., Bolton, D.K., White, J.C., Wulder, M.A., van Lier, O.R., Hermosilla, T., 2021. Modelling lidar-derived estimates of forest attributes over space and time: a review of approaches and future trends. Remote Sens. Environ. 260, 112477.

Davies, A.B., Asner, G.P., 2014. Advances in animal ecology from 3D-LiDAR ecosystem mapping. Trends Ecol. Evol. 29, 681–691.

de Vries, J.P.R., Koma, Z., WallisDeVries, M.F., Kissling, W.D., 2021. Identifying fine-scale habitat preferences of threatened butterflies using airborne laser scanning. Divers. Distrib. 27, 1251–1264.

Dubayah, R., Blair, J.B., Goetz, S., Fatoyinbo, L., Hansen, M., Healey, S., Hofton, M., Hurtt, G., Kellner, J., Luthcke, S., Armston, J., Tang, H., Duncanson, L., Hancock, S., Jantz, P., Marselis, S., Patterson, P.L., Qi, W., Silva, C., 2020. The global ecosystem dynamics investigation: high-resolution laser ranging of the Earth's forests and topography. Sci. Rem. Sens. 1, 100002.

Fagúndez, J., 2012. Heathlands confronting global change: drivers of biodiversity loss from past to future scenarios. Ann. Bot. 111, 151–172.

Fahrig, L., Baudry, J., Brotons, L., Burel, F.G., Crist, T.O., Fuller, R.J., Sirami, C., Siriwardena, G.M., Martin, J.-L., 2011. Functional landscape heterogeneity and animal biodiversity in agricultural landscapes. Ecol. Lett. 14, 101–112.

Fiore, S., Elia, D., Blanquer, I., Brasileiro, F.V., Nuzzo, A., Nassisi, P., Rufino, I.A.A., Seijmonsbergen, A.C., Anders, N.S., de Galvão, O.C., de Cunha, B.L., Caballer, M., Sousa-Baena, M.S., Canhos, V.P., Aloisio, G., 2017. BioClimate: a science gateway for climate change and biodiversity research in the EUBrazilCloudConnect project. Futur. Gener. Comput. Syst. 94, 895–909.

Gobakken, T., Næsset, E., 2008. Assessing effects of laser point density, ground sampling intensity, and field sample plot size on biophysical stand properties derived from airborne laser scanner data. Can. J. For. Res. 38, 1095–1109.

Goodbody, T.R.H., Coops, N.C., Luther, J.E., Tompalski, P., Mulverhill, C., Frizzle, C., Fournier, R., Furze, S., Herniman, S., 2021. Airborne laser scanning for quantifying criteria and indicators of sustainable forest management in Canada. Can. J. For. Res. 51, 972–985.

Haddad, N.M., Brudvig, L.A., Clobert, J., Davies, K.F., Gonzalez, A., Holt, R.D., Lovejoy, T.E., Sexton, J.O., Austin, M.P., Collins, C.D., Cook, W.M., Damschen, E.I., Ewers, R.M., Foster, B.L., Jenkins, C.N., King, A.J., Laurance, W.F., Levey, D.J., Margules, C.R., Melbourne, B.A., Nicholls, A.O., Orrock, J.L., Song, D.-X., Townshend, J.R., 2015. Habitat fragmentation and its lasting impact on Earth's ecosystems. Sci. Adv. 1, e1500052.

Höfle, B., Hollaus, M., Hagenauer, J., 2012. Urban vegetation detection using radiometrically calibrated small-footprint full-waveform airborne LiDAR data. ISPRS J. Photogramm. Remote Sens. 67, 134–147.

Hyyppä, J., Hyyppä, H., Leckie, D., Gougeon, F., Yu, X., Maltamo, M., 2008. Review of methods of small-footprint airborne laser scanning for extracting forest inventory data in boreal forests. Int. J. Remote Sens. 29, 1339–1366.

Inomata, T., Triadan, D., Vázquez López, V.A., Fernandez-Diaz, J.C., Omori, T., Méndez Bauer, M.B., García Hernández, M., Beach, T., Cagnato, C., Aoyama, K., Nasu, H., 2020. Monumental architecture at Aguada Fénix and the rise of Maya civilization. Nature 582, 530–533.

Jakubowski, M.K., Guo, Q., Kelly, M., 2013. Tradeoffs between lidar pulse density and forest measurement accuracy. Remote Sens. Environ. 130, 245–253.

Kissling, W.D., Seijmonsbergen, A., Foppen, R., Bouten, W., 2017. eEcoLiDAR, eScience infrastructure for ecological applications of LiDAR point clouds: reconstructing the 3D ecosystem structure for animals at regional to continental scales. Res. Ideas Outcomes 3, e14939.

Kissling, W.D., Shi, Y., Koma, Z., Meijer, C., Ku, O., Nattino, F., Seijmonsbergen, A.C., Grootes, M.W., 2022. Country-wide data of ecosystem structure from the third Dutch airborne laser scanning survey. Data in Brief submitted.

Koma, Z., Grootes, M.W., Meijer, C.W., Nattino, F., Seijmonsbergen, A.C., Sierdsema, H., Foppen, R., Kissling, W.D., 2021a. Niche separation of wetland birds revealed from airborne laser scanning. Ecography 44, 907–918.

Koma, Z., Seijmonsbergen, A.C., Kissling, W.D., 2021b. Classifying wetland-related land cover types and habitats using fine-scale lidar metrics derived from country-wide airborne laser scanning. Rem. Sens. Ecol. Conserv. 7, 80–96.

Koma, Z., Zlinszky, A., Bekő, L., Burai, P., Seijmonsbergen, A.C., Kissling, W.D., 2021c. Quantifying 3D vegetation structure in wetlands using differently measured airborne laser scanning data. Ecol. Indic. 127, 107752.

Kraus, K., Pfeifer, N., 1998. Determination of terrain models in wooded areas with airborne laser scanner data. ISPRS J. Photogramm. Remote Sens. 53, 193–203.

Lefsky, M.A., Cohen, W.B., Parker, G.G., Harding, D.J., 2002. Lidar remote sensing for ecosystem studies. Bioscience 52, 19–30.

Luo, S., Wang, C., Pan, F., Xi, X., Li, G., Nie, S., Xia, S., 2015. Estimation of wetland vegetation height and leaf area index using airborne laser scanning data. Ecol. Indic. 48, 550–559.

MacArthur, R., MacArthur, J.W., 1961. On bird species diversity. Ecology 42, 594–598.

Maltamo, M., Naesset, E., Vauhkonen, J., 2014. Forestry Applications of Airborne Laser Scanning - Concepts and Case Studies. Springer, Dordrecht.

Matasci, G., Hermosilla, T., Wulder, M.A., White, J.C., Coops, N.C., Hobart, G.W., Bolton, D.K., Tompalski, P., Bater, C.W., 2018. Three decades of forest structural dynamics over Canada's forested ecosystems using Landsat time-series and lidar plots. Remote Sens. Environ. 216, 697–714.

Meijer, C., Grootes, M.W., Koma, Z., Dzigan, Y., Gonçalves, R., Andela, B., van den Oord, G., Ranguelova, E., Renaud, N., Kissling, W.D., 2020. Laserchicken—a tool for distributed feature calculation from massive LiDAR point cloud datasets. SoftwareX 12, 100626.

Moeslund, J.E., Zlinszky, A., Ejrnæs, R., Brunbjerg, A.K., Bøcher, P.K., Svenning, J.-C., Normand, S., 2019. Light detection and ranging explains diversity of plants, fungi, lichens and bryophytes across multiple habitats and large geographic extent. Ecol. Appl. 29, e01907.

PDAL Contributors, 2020. PDAL Point Data Abstraction Library. Available at. https://zenodo.org/record/2556738#.YlfXxNNByUl. Accessed May 2022.

Peduzzi, A., Wynne, R.H., Fox, T.R., Nelson, R.F., Thomas, V.A., 2012. Estimating leaf area index in intensively managed pine plantations using airborne laser scanner data. For. Ecol. Manag. 270, 54–65.

Pereira, H.M., Ferrier, S., Walters, M., Geller, G.N., Jongman, R.H.G., Scholes, R.J., Bruford, M.W., Brummitt, N., Butchart, S.H.M., Cardoso, A.C., Coops, N.C., Dulloo, E., Faith, D.P., Freyhof, J., Gregory, R.D., Heip, C., Höft, R., Hurtt, G., Jetz, W., Karp, D.S., McGeoch, M.A., Obura, D., Onoda, Y., Pettorelli, N., Reyers, B., Sayre, R., Scharlemann, J.P.W., Stuart, S.N., Turak, E., Walpole, M., Wegmann, M., 2013. Essential biodiversity variables. Science 339, 277–278.

Pereira, H.M., Junker, J., Fernández, N., Maes, J., Beja, P., Bonn, A., Breeze, T., Brotons, L., Bruehlheide, H., Buchhorn, M., Capinha, C., Chow, C., Dietrich, K., Dornelas, M., Dubois, G., Fernandez, M., Frenzel, M., Friberg, N., Fritz, S., Georgieva, I., Gobin, A., Guerra, C., Haande, S., Herrando, S., Jandt, U., Kissling, W. D., Kühn, I., Langer, C., Liquete, C., Lyche Solheim, A., Martí, D., Martin, J.G.C., Masur, A., McCallum, I., Mjelde, M., Moe, J., Moersberger, H., Morán-Ordóñez, A., Moreira, F., Musche, M., Navarro, L.M., Orgiazzi, A., Patchett, R., Penev, L., Pino, J., Popova, G., Potts, S., Ramon, A., Sandin, L., Santana, J., Sapundzhieva, A., See, L., Shamoun-Baranes, J., Smets, B., Stoev, P., Tedersoo, L., Tiimann, L., Valdez, J., Vallecillo, S., Van Grunsven, R.H.A., Van De Kerchove, R., Villero, D., Visconti, P., Weinhold, C., Zuleger, A.M., 2022. Europa biodiversity observation network: integrating data streams to support policy. ARPHA Preprints 3.

Pettorelli, N., Wegmann, M., Skidmore, A., Mücher, S., Dawson, T.P., Fernandez, M., Lucas, R., Schaepman, M.E., Wang, T., O'Connor, B., Jongman, R.H.G., Kempeneers, P., Sonnenschein, R., Leidner, A.K., Böhm, M., He, K.S., Nagendra, H., Dubois, G., Fatoyinbo, T., Hansen, M.C., Paganini, M., de Klerk, H.M., Asner, G.P., Kerr, J.T., Estes, A.B., Schmeller, D.S., Heiden, U., Rocchini, D., Pereira, H.M., Turak, E., Fernandez, N., Lausch, A., Cho, M.A., Alcaraz-Segura, D., McGeoch, M.A., Turner, W., Mueller, A., St-Louis, V., Penner, J., Vihervaara, P., Belward, A., Reyers, B., Geller, G.N., 2016. Framing the concept of satellite remote sensing essential biodiversity variables: challenges and future directions. Rem. Sens. Ecol. Conserv. 2, 122–131.

Pfeifer, N., Mandlburger, G., Otepka, J., Karel, W., 2014. OPALS – a framework for airborne laser scanning data analysis. Comput. Environ. Urban. Syst. 45, 125–136.

Provoost, S., Jones, M.L.M., Edmondson, S.E., 2011. Changes in landscape and vegetation of coastal dunes in Northwest Europe: a review. J. Coast. Conserv. 15, 207–226.

Rapidlasso GmbH, 2022. LAStools. Available at. https://rapidlasso.com/LAStools/. Accessed 7 April 2022.

Rocklin, M., 2015. Dask: Parallel computation with blocked algorithms and task scheduling. In: Huff, K., Bergstra, J. (Eds.), Proceedings of the 14th Python in Science Conference, pp. 130–136.
Roth, R.R., 1976. Spatial heterogeneity and bird species diversity. Ecology 57, 773–782.
Roussel, J.-R., Auty, D., Coops, N.C., Tompalski, P., Goodbody, T.R.H., Meador, A.S., Bourdon, J.-F., de Boissieu, F., Achim, A., 2020. lidR: an R package for analysis of airborne laser scanning (ALS) data. Remote Sens. Environ. 251, 112061.
Skidmore, A.K., Pettorelli, N., Coops, N.C., Geller, G.N., Hansen, M., Lucas, R., Mücher, C.A., O'Connor, B., Paganini, M., Pereira, H.M., Schaepman, M.E., Turner, W., Wang, T., Wegmann, M., 2015. Agree on biodiversity metrics to track from space. Nature 523, 403–405.
Stereńczak, K., Laurin, G.V., Chirici, G., Coomes, D.A., Dalponte, M., Latifi, H., Puletti, N., 2020. Global airborne laser scanning data providers database (GlobALS)—a new tool for monitoring ecosystems and biodiversity. Remote Sens. 12, 1877.
Tews, J., Brose, U., Grimm, V., Tielbörger, K., Wichmann, M.C., Schwager, M., Jeltsch, F., 2004. Animal species diversity driven by habitat heterogeneity/diversity: the importance of keystone structures. J. Biogeogr. 31, 79–92.
Valbuena, R., O'Connor, B., Zellweger, F., Simonson, W., Vihervaara, P., Maltamo, M., Silva, C.A., Almeida, D.R.A., Danks, F., Morsdorf, F., Chirici, G., Lucas, R., Coomes, D.A., Coops, N.C., 2020. Standardizing ecosystem morphological traits from 3D information sources. Trends Ecol. Evol. 35, 656–667.
Vierling, K.T., Vierling, L.A., Gould, W.A., Martinuzzi, S., Clawges, R.M., 2008. Lidar: shedding new light on habitat characterization and modeling. Front. Ecol. Environ. 6, 90–98.
Vihervaara, P., Auvinen, A.-P., Mononen, L., Törmä, M., Ahlroth, P., Anttila, S., Böttcher, K., Forsius, M., Heino, J., Heliölä, J., Koskelainen, M., Kuussaari, M., Meissner, K., Ojala, O., Tuominen, S., Viitasalo, M., Virkkala, R., 2017. How essential biodiversity variables and remote sensing can help national biodiversity monitoring. Glob. Ecol. Conserv. 10, 43–59.
Wang, Y., Koulouzis, S., Bianchi, R., Li, N., Shi, Y., Timmermans, J., Kissling, W.D., Zhao, Z., 2022. Scaling notebooks as re-configurable cloud workflows. Data Intellig. 4, 409–425.
Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., 't Hoen, P.A.C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B., 2016. The FAIR guiding principles for scientific data management and stewardship. Sci. Data 3, 160018.
Wulder, M.A., White, J.C., Nelson, R.F., Næsset, E., Ørka, H.O., Coops, N.C., Hilker, T., Bater, C.W., Gobakken, T., 2012. Lidar sampling for large-area forest characterization: a review. Remote Sens. Environ. 121, 196–209.
Yu, X., Litkey, P., Hyyppä, J., Holopainen, M., Vastaranta, M., 2014. Assessment of low density full-waveform airborne laser scanning for individual tree detection and tree species classification. Forests 5, 1011.
Zellweger, F., De Frenne, P., Lenoir, J., Rocchini, D., Coomes, D., 2019. Advances in microclimate ecology arising from remote sensing. Trends Ecol. Evol. 34, 327–341.
Zhao, Z., Koulouzis, S., Bianchi, R., Farshidi, S., Shi, Z., Xin, R., Wang, Y., Li, N., Shi, Y., Timmermans, J., Kissling, W.D., 2022. Notebook-as-a-VRE (NaaVRE): from private notebooks to a collaborative cloud virtual research environment. Software: Practice and Experience 52, 1947–1966.
Zlinszky, A., Mücke, W., Lehner, H., Briese, C., Pfeifer, N., 2012. Categorizing wetland vegetation by airborne laser scanning on Lake Balaton and Kis-Balaton, Hungary. Remote Sens. 4, 1617–1650.