



SMAPVEX19-22 Massachusetts Plant Area Index, Version 1

USER GUIDE

How to Cite These Data

As a condition of using these data, you must include a citation:

Kraatz, S., Cosh, M. H., Kelly, V., Colliander, A., and P. R. Siqueira. 2025. *SMAPVEX19-22 Massachusetts Plant Area Index, Version 1*. [Indicate subset used]. Boulder, Colorado USA. NASA National Snow and Ice Data Center Distributed Active Archive Center. <https://doi.org/10.5067/14BX4UOPEVRX>. [Date Accessed].

FOR QUESTIONS ABOUT THESE DATA, CONTACT NSIDC@NSIDC.ORG

FOR CURRENT INFORMATION, VISIT https://nsidc.org/data/SV19MA_PA



National Snow and Ice Data Center

TABLE OF CONTENTS

1	DATA DESCRIPTION	2
1.1	Summary	2
1.2	Parameters	2
1.3	File Information	2
1.3.1	Format	2
1.3.2	File Contents	2
1.3.3	Naming Convention	3
1.3.4	Directory Structure	4
1.4	Spatial Information	6
1.4.1	Coverage	6
1.4.2	Resolution	8
1.4.3	Geolocation	9
1.5	Temporal Information	10
1.5.1	Coverage and Resolution	10
2	DATA ACQUISITION AND PROCESSING	10
2.1	Acquisition and Processing	10
2.1.1	Pre-processing Steps	11
2.1.2	Processing Steps	11
2.1.3	Post-processing Steps	12
3	VERSION HISTORY	13
4	RELATED DATA SETS	13
5	RELATED WEBSITES	13
6	REFERENCES	14
7	ACKNOWLEDGMENTS	14
8	DOCUMENT INFORMATION	14
8.1	Publication Date	14
8.2	Date Last Updated	14

1 DATA DESCRIPTION

1.1 Summary

These data consist of plant area index (PAI) and the RGB images from which the PAI values were derived. Data image files were taken between April 2019 to December 2022 in the vicinity of Petersham, Massachusetts during the SMAPVEX19-22 (Soil Moisture Active Passive Validation Experiment 2019-2022) field campaign. The location was chosen due to its forested land cover, as SMAPVEX19-22 aims to validate satellite derived soil moisture estimates in forested areas.

1.2 Parameters

This data set includes plant area index (PAI) values derived from each data image file.

1.3 File Information

1.3.1 Format

Images are provided as joint photographic experts group (.jpg or .jpeg) files. PAI values, where each value is derived from each image taken, are compiled into multiple single comma-separated value (.csv) files.

1.3.2 File Contents

PAI values and the multiple variables used to process PAI are presented in .csv files, generated from image analysis in EzPAI software—a free open-source Python-based tool that automates the PAI-processing workflow and data quality screening. Specific variable names and their descriptions are provided in Table 1.

Table 1. Variable Names and Descriptions

Variable Name	Description
timestamp	Acquisition date, specified as 4-digit year, 2-digit month, 2-digit day, 2-digit hour, 2-digit minute, 2-digit seconds; for example, 2019-04-23 14:40:30
Name	File name helping to uniquely identify each photo
lmb	Bin location for the canopy; used for bilevel thresholding in the software
lmc	Bin counts for the canopy; used for bilevel thresholding in the software
rm	Bin location counts for sky; used for bilevel thresholding in the software

Variable Name	Description
rmxc	Bin counts for sky; used for bilevel thresholding in the software
rb_l	Bin locations towards left of histogram resulting from corner detection for sky (rb _l)
rb_r	Bin locations towards right of histogram resulting from corner detection for plant (rb _r)
sky	Value for sky blueness index (BSI) of the sky; > 0.54 means clear sky
minpixarea	The value of the smallest large gap pixel counts (LGC) expressed as percentage of the image
GF	Gap friction
QC	Image quality control (QC); alternatively, also referred to as data quality (DQ). QC is either standard (0) or impaired (1 or 2). See also: <i>Post-processing Steps</i> . When QC = 0, workflow for thresholding and calculating PAI follows the standard in existing literature. When QC = 1, first alternative processing pathway, when the search for the right maximum bin was found at bin <160 but did not have a greater count than the last non-zero bin. When QC = 2, second alternative processing pathway, when the right maximum search did not find any result by the time it reached bin 128.
delta	The difference between Rosin bin locations; specifically, (rb _l - rb _r)
CC	Crown Cover (fraction of pixels that do not lie in between-crown gaps)
CP	Crown Porosity (0 to 1, the gap fraction within crowns)
PAI	Plant Area Index (not calibrated to in situ, assumed k = 0.65)
CI	Clumping Index (0 to 1, lower is more clumped)

1.3.3 Naming Convention

1.3.3.1 Data File Name with PAI values

The data file name containing PAI values of each image taken are provided in .csv files with the following convention as described in detail on Table 2.

SV19MA_PAI_[cameraID]_[YYYYMMDD]_V01.0.csv

An example data file name is:

SV19MA_PAI_416cam_20190509_V01.0.csv

Table 2. File naming convention for data files

Variable	Description
SV19MA	Indicates SMAP Validation Experiment took place from 2019 to 2022 in the vicinity of Petersham, Massachusetts
[cameraID]	ID camera number (designated as three digits followed by "cam", with first digit starting as 4 for photos taken in Petersham, Massachusetts)
[YYYYMMDD]	Start date within .csv file representing the earliest photograph listed within that particular subfolder , designated in the following format: 4-digit year, 2-digit month, 2-digit day (YYYYMMDD); See also: <i>Directory Structure</i>
V01.0	Indicates version 01.0
.csv	Indicates file extension as comma-separated value (.csv)

1.3.3.2 Image File Name

The image data files follow the following convention with details described in Table 3.

SV19MA_PAI_[cameraID]_[photoID]_YYYYMMDDhhmmEST_V01.0.jpg

An example file name is:

SV19MA_PAI_401cam_WSCT3077_20200529110001EST_V01.0.jpg

Table 3. File naming convention for image files

Variable	Description
SV19MA	Indicates SMAP Validation Experiment took place from 2019 to 2022 in the vicinity of Petersham, Massachusetts
PAI	Indicates plant area index (PAI)
[cameraID]	ID camera number (designated as three digits followed by "cam", with first digit starting as 4 for photos taken in Petersham, Massachusetts)
[photoID]	Photo ID details
YYYYMMDDhhmmssEST	Acquisition date and time, specified as 4-digit year (YYYY), 2-digit month (MM), 2-digit day (DD), 2-digit hour (hh), 2-digit minute (mm) in Eastern Time Zone
V01.0	Indicates version 01.0
.jpg	Indicates file extension as joint photographic experts group (.jpg)

1.3.4 Directory Structure

Data .csv files are placed within each subfolder of granules based on their respective start date, that is, the earliest photograph listed within that particular .csv file. Table 4 describes in detail how to find .csv metadata on a particular photo from a particular camera.

For example, to find .csv metadata, including PAI values, for the following .jpg:

SV19MA_PAI_412cam_WSCT8787_20220805160001EST_V01.0.jpg,

You will need to access the following CSV file:

SV19MA_PAI_412cam_20190508.csv,

And search for the corresponding photograph name; in this case, WSCT8787.

Table 4. Key for accessing metadata from a particular image

Camera number	Start date (Earliest photograph listed within the CSV file)	Corresponding CSV file name
401	2019-04-23	SV19MA_PAI_401cam_20190423.csv
403	2019-04-23	SV19MA_PAI_403cam_20190423.csv
404	2019-04-23	SV19MA_PAI_404cam_20190423.csv
406	2019-04-23	SV19MA_PAI_406cam_20190423.csv
407	2019-04-23	SV19MA_PAI_407cam_20190423.csv
408	2019-04-24	SV19MA_PAI_408cam_20190424.csv
409	2019-04-24	SV19MA_PAI_409cam_20190424.csv
411	2019-04-24	SV19MA_PAI_411cam_20190424.csv
412	2019-05-08	SV19MA_PAI_412cam_20190508.csv
413	2019-05-08	SV19MA_PAI_413cam_20190508.csv
414	2019-05-08	SV19MA_PAI_414cam_20190508.csv
415	2019-05-08	SV19MA_PAI_415cam_20190508.csv
416	2019-05-09	SV19MA_PAI_416cam_20190509.csv
417	2019-05-09	SV19MA_PAI_417cam_20190509.csv
418	2019-05-09	SV19MA_PAI_418cam_20190509.csv
419	2020-06-26	SV19MA_PAI_419cam_20200626.csv
420	2019-05-09	SV19MA_PAI_420cam_20190509.csv
421	2019-05-09	SV19MA_PAI_421cam_20190509.csv
422	2019-05-09	SV19MA_PAI_422cam_20190509.csv
423	2019-05-09	SV19MA_PAI_423cam_20190509.csv
424	2019-04-23	SV19MA_PAI_424cam_20190423.csv

1.4 Spatial Information

1.4.1 Coverage

The Petersham, MA site is defined by the following geographic coordinates:

Northernmost Latitude: 42.72° N

Southernmost Latitude: 42.32° N

Easternmost Longitude: 71.91° W

Westernmost Longitude: 72.33° W

Figure 1 displays the field locations, sampling sites, and the dominant land use types in both the SV19MA_PA I and SV19MB_PA I study regions.

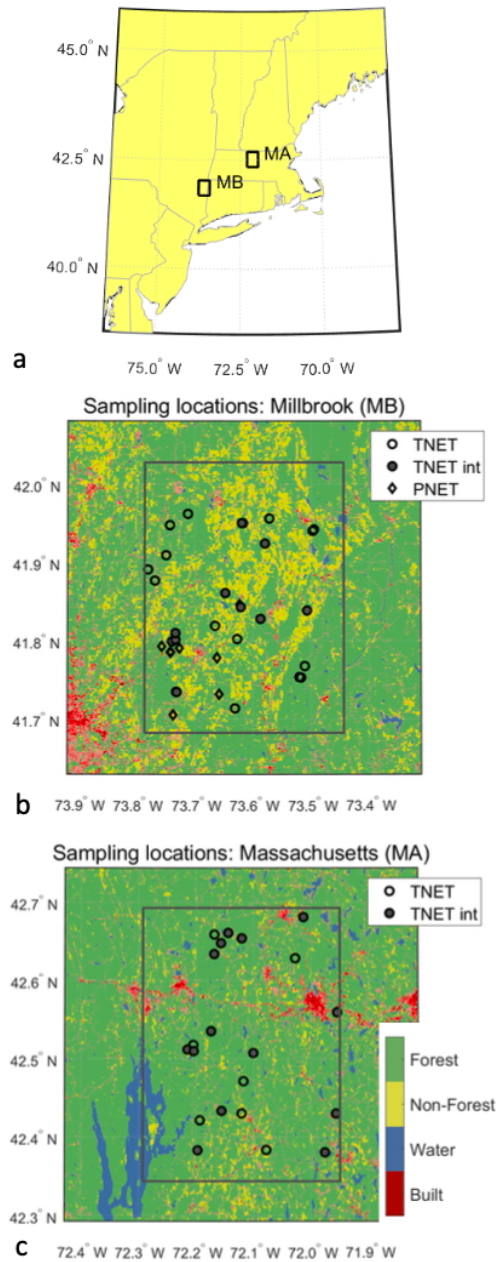


Figure 1. SMAPVEX19-22 Sampling Locations. (a) Map of the experiment locations near Amherst, Massachusetts (MA) and Millbrook, New York (MB). SV19MA_PAI data comes from the Petersham, MA site. (b) Locations of the temporary 25 soil moisture stations (TNET) in the Millbrook pixel and the locations of the eight long-term stations (PNET) operated in the area. “TNET int” stands for the intensive sites. (c) Locations of the temporary 23 soil moisture stations in the Massachusetts pixel. The background colors in (b) and (c) are based on the National Land Cover Database. The rectangles depict the SMAP pixel, and the outlines of (b) and (c) are 50 km by 50 km (Colliander et al., 2025).

1.4.2 Resolution

These data are point measurements, where the field of view contains a canopy of several trees. These various instantaneous point measurements were then processed to calculate plant area index (PAI) of each image. Figures 2 and 3 display the sampling techniques and spatial layout strategy of the SMAPVEX19-22 campaign.

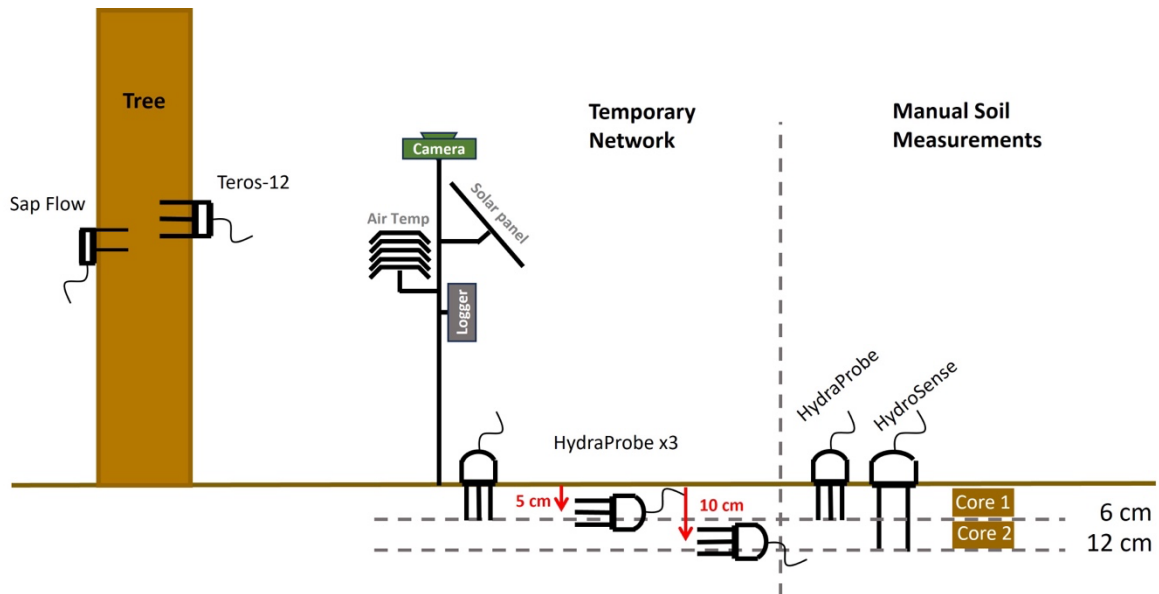


Figure 2. Diagram of the temporary network measurements and manual soil moisture measurements. The temporary networks included soil moisture measurements at three depths, which also include temperature measurements, air temperature, zenith camera, and dielectric (Teros-12) and sap flow sensors in tree at a subset of locations. The manual soil moisture measurements included a triplet of HydraProbe and HydroSense measurements at 14 locations around the stations, and core samples obtained from two depths at four locations around the stations (see also Figure 3) (Colliander et al., 2025).

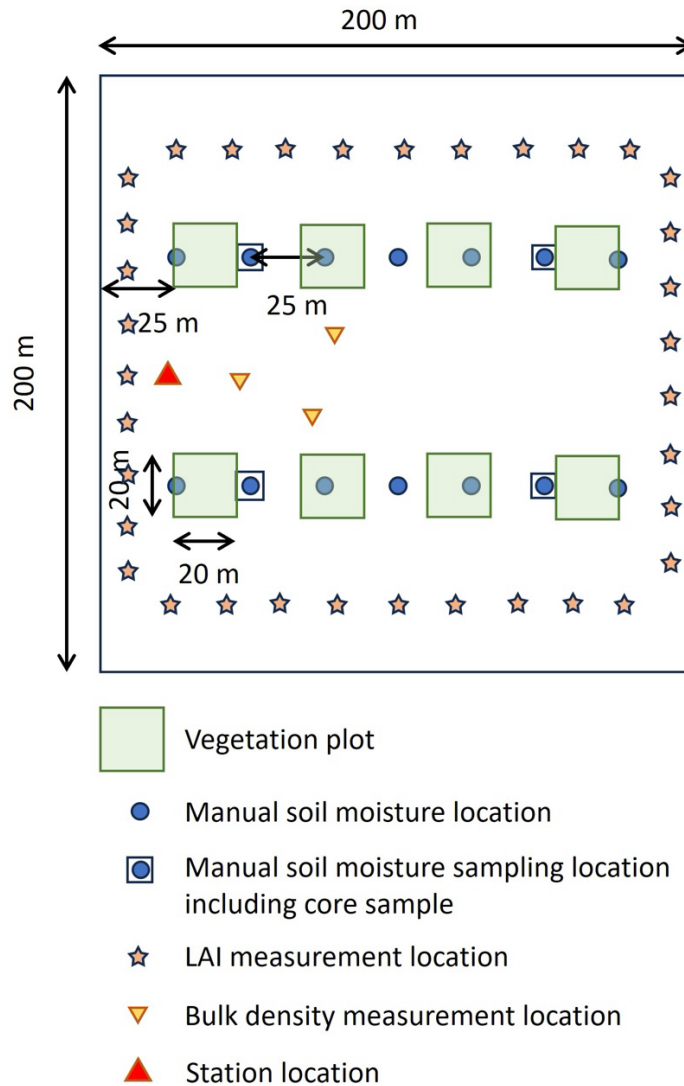


Figure 3. Layout of the soil moisture and vegetation transects. The sampling was intended to represent the predominant hydrological and ecological features of the 200 m by 200 m plot. Soil moisture was measured along two transects extending across the plot (Colliander et al., 2025).

1.4.3 Geolocation

The following table provides information for geolocating this data set.

Table 5. Geolocation Details

Geographic coordinate system	WGS84 - World Geodetic System 1984, used in GPS
Projected coordinate system	N/A
Longitude of true origin	0°
Latitude of true origin	N/A

Scale factor at longitude of true origin	N/A
Datum	WGS_1984
Ellipsoid/spheroid	WGS 84
Units	degree
False easting	N/A
False northing	N/A
EPSG code	4326
PROJ4 string	+proj=longlat +datum=WGS84 +no_defs +type=crs
Reference	https://epsg.org/crs_4326/WGS-84.html

1.5 Temporal Information

1.5.1 Coverage and Resolution

This data set spans 01 Apr 2019 to 01 Dec 2022. Individual data image files contain specific sampling date and time. PAI is provided within multiple comma-separated values (.csv) data files.

2 DATA ACQUISITION AND PROCESSING

2.1 Acquisition and Processing

As part of the SMAPVEX19-22 (Soil Moisture Active Passive Validation Experiment 2019-2022) field campaign, RGB images of the tree canopy were captured and processed to calculate plant area index (PAI) of each image. Digital cover photography (DCP) is suitable for repeat PAI retrievals over long periods at lower costs due to minimal maintenance and minimal personnel. Through DCP, PAI values were calculated as an output product from the EzPAI workflow software, a free open-source research code written in Python that enables processing workflow and automated data quality screening.

The basic workflow is as follows: First, raw image files are input into EzPAI software where the data goes through pre-processing; this involves running the hour screening script and blur detection script. Next, the processing involves data filtering thresholds, specifically plant-sky thresholding and identifying any large gaps in the data. Once these steps are completed, PAI is calculated with post-processing bias corrections, which include data quality screening and correcting for bias between cloud versus clear sky. Finally, post-processed results are compiled to produce a .csv file with the PAI values listed for each image that was input into the system.

2.1.1 Pre-processing Steps

The hour screening script accounts for how illumination conditions change over the year using the following scheme: (1) if the months were April, May, August, or September, it only allowed images with timestamps between 7 AM and 6 PM local EST; (2) if the months were June or July, it only allowed images with timestamps between 6 AM and 7 PM local Eastern time (ET); (3) if the months were October or March, it only allowed images with timestamps between 8 AM and 5 PM local ET; and (4) if the months were November or February, it only allowed images with timestamps between 9 AM and 4 PM local ET. Note, for this SMAPVEX19-22 data set, the option to screen data by hour of the day to help remove images with sun glare was not used.

The blur screening script involves reading the .csv output from the hour screening pre-processing step. Blur detection is based on a less strict form of the variance of the Laplacian approach, where data were required to have either a variance ≥ 0.01 or maximum ≥ 1.08 after the Laplacian (kernel size = 3) has been applied. To speed up blur detection, input images were scaled to 25% of their original size after omitting the bottom 100 pixels corresponding to the metadata overlay. This step is computationally expensive and takes nearly half the processing time of the DCP processing script. However, the approach successfully identified most poor-quality images (at times, ~90% of a site-year), resulting in overall savings in processing time.

2.1.2 Processing Steps

Plant-sky thresholding is performed by using only the blue channel of each image to delineate the plant and sky, and thereby produce a histogram as an output. Under ideal conditions, the histogram will have two peaks, one towards lower and another towards greater digital numbers (DN), respectively corresponding to unmixed plant and sky pixels. The Rosin corner-detection method was used to identify pure and mixed pixels for plant and sky. Using a search that started from the left and right ends of the histogram, the left and right peaks of the histogram were detected. On each end of the histogram, a line was constructed by extending from the first non-empty bin (at counts=0) to the y-coordinate (counts) of the peak on the right and left, respectively. Then the point of maximum curvature of each peak was determined by identifying the bin which corresponded to the perpendicular line having the greatest distance to a non-empty bin; these were labeled left and right Rosin bins, rb_l and rb_r , respectively. The number of bins between rb_l and rb_r (i.e., $rb_l - rb_r$) then identifies image DNs that are mixed pixels. This range can then be proportioned into a single binary plant vs. sky delineating threshold (b_{thr} , Eq. 1), as needed.

$$b_{thr} = rb_r + \text{int}(f_{part} * (rb_l - rb_r)), \quad \text{Eq. 1}$$

Where partitioning factors, f_{part} , may have different values according to whether an image was categorized as clear or cloudy.

To iteratively identify any large gaps of data during the DCP processing, the contour finding approach `CV2.findContours()` from OpenCV software was employed. The input is the binary result of the plant and sky classification ($DN = 0$ and 1 , respectively) that resulted from using the b_{thr} value calculated by Eq. 1. All found contours were then screened according to a contour size threshold (FCVAL in EzPAI script) of 10000. For each contour that exceeds FCVAL, the intersecting image pixels are assigned a DN of 255 and summated for each image to give the large gap pixel counts (LGC). The FCVAL threshold was estimated from visual inspection to yield reasonable values for the minimum size of the large between-crown pixel gaps relative to the image size. By setting an FCVAL value of 10000, LGC corresponded to about $> 0.3\%$ of the image.

2.1.3 Post-processing Steps

Although the plant-sky thresholding step worked well for most images (indicated by a data quality (DQ) flag of 0), additional data quality assignments were created because the two-threshold plant vs. sky approach would sometimes yield poor results or not find a solution (i.e., $rb_l = rb_r$). Furthermore, because the pre-processing was not overly strict, some poor-quality images (i.e., blur, glare) were also processed. Some site years contained many poor-quality images, likely due to biofilm or other weathering of the plastic lens covers. However, visual inspection of many of these images indicated that it was still possible to obtain reasonable PAI estimates.

To facilitate PAI retrievals for cases when the standard workflow (DQ=0) either did not obtain a reasonable estimate or no estimate, two additional flag conditions for different processing pathways were created. In the first condition, DQ=1 was assigned when the search for the right maximum bin was found at bin < 160 but did not have a greater count than the last non-zero bin. In the second condition, DQ=2 was assigned for when the right maximum bin search did not find any result by the time it reached bin 128. Adding these two options to the binary thresholding enabled considerably more PAI values to be estimated which was critical for some site-years.

Tracking sky conditions to quantify and correct for any PAI biases is an important factor in PAI image processing. For determination of clear or cloudy sky conditions, EzPAI uses the sky blueness index (BSI) by utilizing the red, green, and blue channels at each pixel (Eq. 2).

$$BSI = DN_{blue} / (DN_{red} + DN_{green}) \quad \text{Eq. 2}$$

Based on the DCP data and visual inspection, a cloudy vs. clear threshold was set at 0.54 (*cloudythr* in the script), with $BSI < 0.54$ corresponding to cloudy.

3 VERSION HISTORY

Table 6. Version History

Version	Date Implemented	Impacted Temporal Coverage	Description of Changes
v01.0	October 2025	01 Apr 2019 to 01 Dec 2022	Initial release

4 RELATED DATA SETS

[SMAPVEX19-22 Millbrook Plant Area Index, Version 1](#)

[SMAPVEX19-22 Massachusetts Airborne Lidar, Version 1](#)

[SMAPVEX19-22 Massachusetts Lidar Derived Digital Elevation Model, Version 1](#)

[SMAPVEX19-22 Massachusetts Lidar Derived Digital Surface Model, Version 1](#)

[SMAPVEX19-22 Massachusetts Temporary Soil Moisture Network, Version 1](#)

[SMAPVEX19-22 Massachusetts Vegetation Optical Depth, Version 1](#)

[SMAPVEX19-22 Millbrook Airborne Lidar, Version 1](#)

[SMAPVEX19-22 Millbrook Lidar Derived Digital Elevation Model, Version 1](#)

[SMAPVEX19-22 Millbrook Lidar Derived Digital Surface Model, Version 1](#)

[SMAPVEX19-22 Millbrook Temporary Soil Moisture Network, Version 1](#)

[SMAPVEX19-22 Massachusetts UAVSAR Mosaics, Version 1](#)

[SMAPVEX19-22 Millbrook UAVSAR Mosaics, Version 1](#)

5 RELATED WEBSITES

[SMAP Validation Data](#)

[SMAP Overview](#)

6 REFERENCES

Colliander, A., Cosh, M. H., Bourgeau-Chavez, L., Kelly, V., Kraatz, S., Siqueira, P., Walker, V. A., Chen, X., Roy, A., Lakhankar, T., McDonald, K., Steiner, N., Kurum, M., Kim, S., Berg, A., Xu, X., Misra, S., Ogut, M., Vittucci, C., ... Yueh, S. H. (2025). SMAP validation experiment 2019-2022 (SMAPVEX19-22): Field campaign to improve soil moisture and vegetation optical depth retrievals in temperate forests. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 1–23.
<https://doi.org/10.1109/JSTARS.2025.3553085>

7 ACKNOWLEDGMENTS

This research used resources provided by the SCINet project of the USDA Agricultural Research Service, ARS project number 0500-00093-001-00-D.

8 DOCUMENT INFORMATION

8.1 Publication Date

October 2025

8.2 Date Last Updated

October 2025