OPEN SCIENCE, KNOWLEDGE SHARING AND REPRODUCIBILITY AS DRIVERS FOR THE ADOPTION OF FOSS4G IN ENVIRONMENTAL RESEARCH

Ionuț Iosifescu Enescu[1], Gian-Kasper Plattner[1], Leo Bont, Marielle Fraefel[1], Rolf Meile[1], Thomas Kramer[1], Lucia Espona-Pernas[1], Dominik Haas-Artho[1], Martin Hägeli[1], Konrad Steffen[1]

[1] Swiss Federal Institute for Forest, Snow and Landscape Research WSL, Birmensdorf, Switzerland
(ionut.iosifescu, gian-kasper.plattner, leo.bont, marielle.fraefel, rolf.meile, thomas.kramer, lucia.espona, dominik.haas, martin.haegeli, konrad.steffen)@wsl.ch

Commission IV, WG IV/4

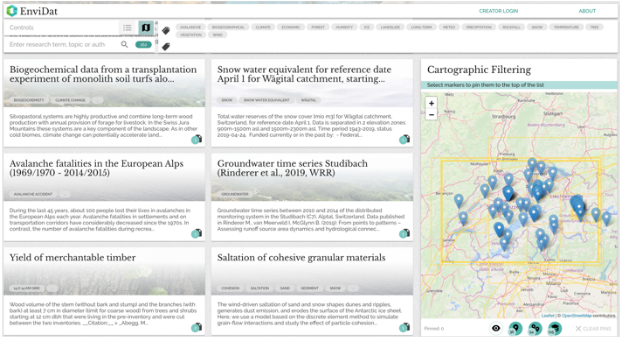
KEY WORDS: Open Science, Knowledge Sharing, Reproducibility, FOSS4G, EnviDat, Environmental Data Repository, Environmental Research, Research Data Publication, Research Data Management

ABSTRACT:

Support for open science is a highly relevant user requirement for the environmental data portal EnviDat. EnviDat, the institutional data portal and publication data repository of the Swiss Federal Research Institute WSL, actively implements the FAIR (Findability, Accessibility, Interoperability and Reusability) principles and provides a range of services in the area of research data management. Open science, with its requirements for improved knowledge sharing and reproducibility, is driving the adoption of free and open source software for geospatial (FOSS4G) in academic research. Open source software can play a key role in the proper documentation of data sets, processes and methodologies, because it supports the transparency of methods and the precise documentation of all steps needed to achieve the published results. EnviDat actively supports these activities to enhance its support for open science. With EnviDat, WSL contributes to the ongoing cultural evolution in research towards open science and opportunities for distant collaboration.

1. INTRODUCTION

EnviDat – www.envidat.ch – is the environmental data portal and repository of the Swiss Federal Institute for Forest, Snow and Landscape Research WSL. EnviDat actively implements the FAIR (Findability, Accessibility, Interoperability and Reusability) principles (Wilkinson at al., 2016) and provides a range of services in the area of research data management that were extensively described in Iosifescu et al. (2018a). With its capability to host and publish research data sets, EnviDat currently provides unified and managed access to more than 160 environmental data sets collected by WSL researchers and their research partners (Figure 1).



EnviDat
Controls
Enter research term, topic or auth
Cartographic Filtering
Select markers to pin them to the top of the list
Pinned: 0

Biogeochemical data from a transplantation experiment of monolith soil turfs alo...

Silvopastoral systems are highly productive and combine long-term wood production with annual provision of forage for livestock. In the Swiss Jura Mountains these systems are a key component of the landscape. As in other cold biomes, climate change can potentially accelerate land...

Snow water equivalent for reference date April 1 for Wägital catchment, starting...

Total water reserves of the snow cover [mio m3] for Wägital catchment, Switzerland, for reference date April 1. Data is separated in 2 elevation zones 900m-1500m asl and 1500m-2300m asl. Time period 1943-2019, status 2019-04-24. Funded currently or in the past by: - Federal...

Avalanche fatalities in the European Alps (1969/1970 - 2014/2015)

During the last 45 years, about 100 people lost their lives in avalanches in the European Alps each year. Avalanche fatalities in settlements and on transportation corridors have considerably decreased since the 1970s. In contrast, the number of avalanche fatalities during recrea...

Groundwater time series Studibach (Rinderer et al., 2019, WRR)

Groundwater time series between 2010 and 2014 of the distributed monitoring system in the Studibach (C7), Alptal, Switzerland. Data published in Rinderer M., van Meerveld I., McGlynn B. (2019): From points to patterns – Assessing runoff source area dynamics and hydrological connec...

Yield of merchantable timber

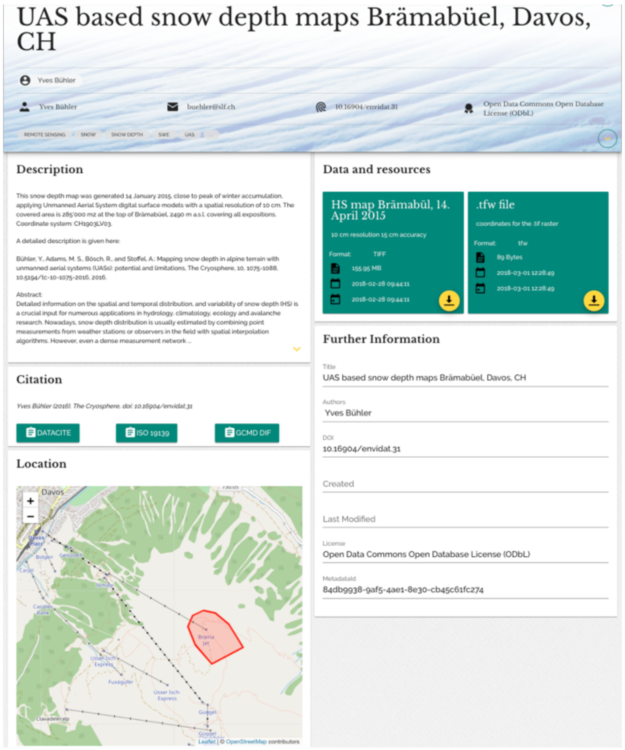
Wood volume of the stem (without bark and stump) and the branches (with bark) at least 7 cm in diameter (limit for coarse wood) from trees and shrubs starting at 12 cm dbh that were living in the pre-inventory and were cut between the two inventories. __Citation:__ > _Abegg, M...

Saltation of cohesive granular materials

The wind-driven saltation of sand and snow shapes dunes and ripples, generates dust emission, and erodes the surface of the Antarctic ice sheet. Here, we use a model based on the discrete element method to simulate grain-flow interactions and study the effect of particle cohesion...

Figure 1. Environmental data sets in EnviDat

The existing datasets, software and additional research information, currently mainly originating from WSL, are accompanied by their documentation with appropriate metadata, including their geospatial footprint. Figure 2 shows an example of a published data set in EnviDat, where the research data is accompanied by proper citation information and Digital Object Identifiers (DOIs).



UAS based snow depth maps Brämabüel, Davos, CH

Yves Bühler

Yves Bühler | buehler@slf.ch | 10.16904/envidat.31 | Open Data Commons Open Database License (ODbL)

Description

This snow depth map was generated 14 January 2015, close to peak of winter accumulation, applying Unmanned Aerial System digital surface models with a spatial resolution of 10 cm. The covered area is 285'000 m2 at the top of Brämabüel, 2490 m a.s.l. covering all expositions. Coordinate system: CH1903LV03.

A detailed description is given here:

Bühler, Y., Adams, M. S., Bösch, R., and Stoffel, A.: Mapping snow depth in alpine terrain with unmanned aerial systems (UASs): potential and limitations, The Cryosphere, 10, 1075-1088, 10.5194/tc-10-1075-2016, 2016.

Abstract:
Detailed information on the spatial and temporal distribution, and variability of snow depth (HS) is a crucial input for numerous applications in hydrology, climatology, ecology and avalanche research. Nowadays, snow depth distribution is usually estimated by combining point measurements from weather stations or observers in the field with spatial interpolation algorithms. However, even a dense measurement network ...

Citation

Yves Bühler (2016). The Cryosphere, doi: 10.16904/envidat.31

Location

Data and resources

HS map Brämabül, 14. April 2015
10 cm resolution 15 cm accuracy
Format: TIFF
155.95 MB
2018-02-28 09:44:11
2018-02-28 09:44:11

.tfw file
coordinates for the .tif raster
Format: tfw
89 Bytes
2018-03-01 12:28:49
2018-03-01 12:28:49

Further Information

Title
UAS based snow depth maps Brämabüel, Davos, CH

Authors
Yves Bühler

DOI
10.16904/envidat.31

Created

Last Modified

License
Open Data Commons Open Database License (ODbL)

MetadataId
84db9938-9af5-4ae1-8e30-cb45c61fc274

Figure 2. Example of a published data set in EnviDat

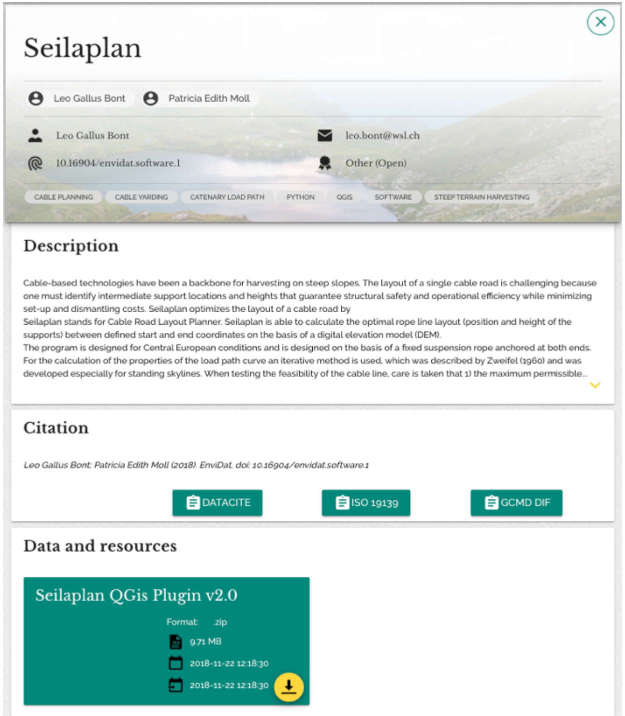
In accordance with the institutional data policy, WSL makes curated, quality-controlled, reusable and publication-ready research data accessible through EnviDat, therefore supporting worldwide data users with an efficient access to various environmental research data sets that were collected by WSL.

2. OPEN SCIENCE SUPPORT IN ENVIDAT

Open science has many aspects and it would be beyond the capabilities of EnviDat to address all aspects comprehensively. We thus focus on one core topics, i.e., knowledge sharing and reproducibility of published research. According to Baker (2016), we have an ongoing "reproducibility crisis", including in the field of Earth and Environment sciences. Unavailable methods or code and missing raw data from original lab are mentioned as major factors for irreproducible results in this study.

Publishing data with formal metadata is often not sufficient to effectively foster open science. For the goal of improving knowledge sharing and reproducibility of published research, the research methods, i.e. computations, need to be made transparent, too. Consequently, the sharing of software is a vital element for the proper understanding of the methodologies presented in research papers.

For this reason, EnviDat encourages WSL scientists to complement data publication with (i) description of research methods and (ii) open source software. An example is provided by Bont et al. (2018) with the opening of a methodology for optimizing the layout of cable roads needed for wood harvesting on steep slopes. It achieves realistic solutions with lower installation costs for cable roads layouts in Central European conditions. The publishing of this methodology is complemented with its implementation in the QGIS plugin "Seilaplan", for everyone to use (Figure 3).



Seilaplan

Leo Gallus Bont · Patricia Edith Moll

Leo Gallus Bont

leo.bont@wsl.ch

10.16904/envidat.software.1

Other (Open)

Description

Cable-based technologies have been a backbone for harvesting on steep slopes. The layout of a single cable road is challenging because one must identify intermediate support locations and heights that guarantee structural safety and operational efficiency while minimizing set-up and dismantling costs. Seilaplan optimizes the layout of a cable road by
Seilaplan stands for Cable Road Layout Planner. Seilaplan is able to calculate the optimal rope line layout (position and height of the supports) between defined start and end coordinates on the basis of a digital elevation model (DEM).
The program is designed for Central European conditions and is designed on the basis of a fixed suspension rope anchored at both ends. For the calculation of the properties of the load path curve an iterative method is used, which was described by Zweifel (1960) and was developed especially for standing skylines. When testing the feasibility of the cable line, care is taken that 1) the maximum permissible...

Citation

Leo Gallus Bont; Patricia Edith Moll (2018). EnviDat. doi: 10.16904/envidat.software.1

Data and resources

Seilaplan QGis Plugin v2.0

Format: .zip

9.71 MB

2018-11-22 12:18:30

2018-11-22 12:18:30

Figure 3. Example of a published FOSS4G in EnviDat

Furthermore, EnviDat also promotes and supports, where possible and practical, the publication of software as Jupyter notebooks. Jupyter notebooks provide a solution for improved documentation and interactive execution of open code in a wide range of programming languages (Python, R, Octave/Matlab, Java or Scala). These programming languages are widely used in environmental research at WSL and well supported by the Jupyter-compatible kernels. In this context, Fraefel (2018) provides one example of a Jupyter notebook to calculate road densities in the neighborhood of sample plot locations using Python (Fraefel, 2018; Iosifescu et al., 2018b).

3. NEW DRIVERS FOR THE ADOPTION OF FOSS4G

Iosifescu et al. (2015) presented several reasons for the increasing adoption of free and open source software for geospatial (FOSS4G) in academic research, such as: the growing stability and maturity of the recent open source software packages, the faster bug-fixing turnover, the increasing availability of professional support, and the flexibility to change and repurpose the open software to tackle new research challenges. These reasons are still valid today, as corroborated by the fact that the EnviDat portal integrates several open software packages such as PostgreSQL/PostGIS, Apache Solr, CKAN and Vue.js in its software architecture.

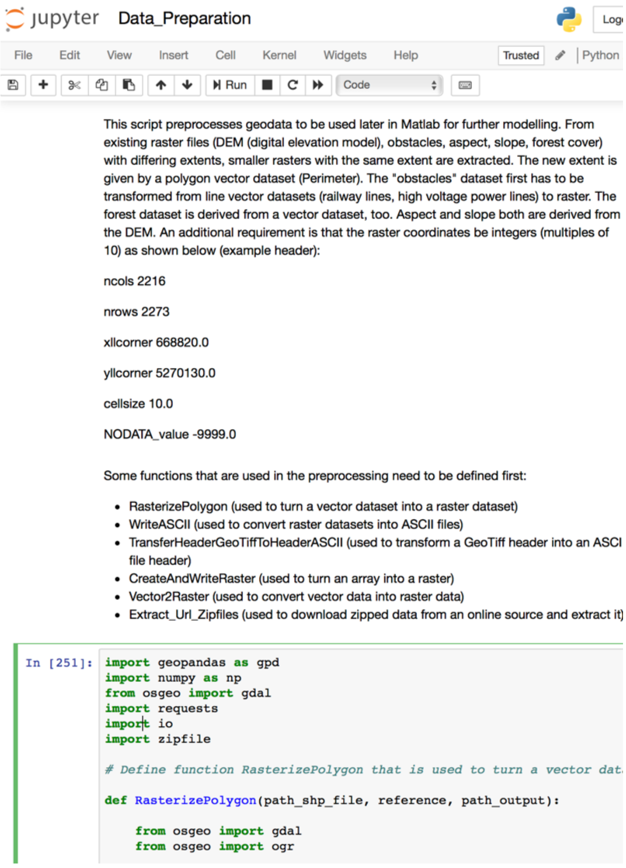
In this contribution, we discuss two novel drivers for the adoption of FOSS4G in environmental research that arose in connection with open science, namely knowledge sharing and reproducibility. We illustrate the larger issue of supporting open science in an environmental data portal such as EnviDat, with the example of Fraefel (2018). The Jupyter notebook for calculating road densities in the neighborhood of sample plot locations was based on the ArcPy site package, allowing the use of ESRI ArcGIS geoprocessing tools in Python. However, although the Python code (through the Jupyter notebook) as well as the corresponding datasets (publicly available in EnviDat) are open, a user would need a proprietary installation with a valid license in order to run it.

We argue that independent research replication at peer-review and after publication is facilitated by (i) the immediate availability of FOSS4G, (ii) the absence of software licensing issues and (iii) the openness of the code even for older versions of a software. Moreover, researchers producing their own code can expect a wider distribution of the produced software, and though it, of their knowledge and expertise. Finally, we would like to note, that these new drivers, at the moment, can be observed in academia. These have not been proven valid for industry, administration or any other domain that is not pursuing academic open science.

4. END TO END REPRODUCIBILITY EXAMPLE FOR ACADEMIC RESEARCH WITH FOSS4G

In EnviDat, open science is supported by the combined publication of bundles of datasets and software as detailed in the next demonstrative use case. The EnviDat record from Bont et al. (2019) demonstrates a fully reproducible preprocessing workflow that generates several output rasters. The notebook presented in Figure 4, extracts harmonized output rasters with the same extent. The extent is given by a polygon vector dataset (Perimeter). These output rasters, such as obstacles, aspect, slope, forest cover, can serve as input data for later computations related to forest accessibility and wood harvesting questions.

The obstacles output is obtained by transforming line vector datasets (railway lines, high voltage power lines) to raster. The forest dataset is derived from a vector dataset. The aspect and slope are both derived from a digital elevation model (DEM). The Python code defines several functions for preprocessing, for example in order to convert raster datasets into ASCII files, to turn an array into a raster or to convert vector data into raster data. The input data is automatically downloaded and extracted from the corresponding EnviDat online public resources. Then, the code will use the GDAL functionality in order to perform the needed geoprocessing, such as resampling of the DEM, computing the slope and aspect from the resampled DEM or generating the obstacle raster. Finally, the computed rasters are written to disk.



This script preprocesses geodata to be used later in Matlab for further modelling. From existing raster files (DEM (digital elevation model), obstacles, aspect, slope, forest cover) with differing extents, smaller rasters with the same extent are extracted. The new extent is given by a polygon vector dataset (Perimeter). The "obstacles" dataset first has to be transformed from line vector datasets (railway lines, high voltage power lines) to raster. The forest dataset is derived from a vector dataset, too. Aspect and slope both are derived from the DEM. An additional requirement is that the raster coordinates be integers (multiples of 10) as shown below (example header):

ncols 2216

nrows 2273

xllcorner 668820.0

yllcorner 5270130.0

cellsize 10.0

NODATA_value -9999.0

Some functions that are used in the preprocessing need to be defined first:

- RasterizePolygon (used to turn a vector dataset into a raster dataset)
- WriteASCII (used to convert raster datasets into ASCII files)
- TransferHeaderGeoTiffToHeaderASCII (used to transform a GeoTiff header into an ASCI file header)
- CreateAndWriteRaster (used to turn an array into a raster)
- Vector2Raster (used to convert vector data into raster data)
- Extract_Url_Zipfiles (used to download zipped data from an online source and extract it)

```
In [251]: import geopandas as gpd
          import numpy as np
          from osgeo import gdal
          import requests
          import io
          import zipfile

          # Define function RasterizePolygon that is used to turn a vector dat

          def RasterizePolygon(path_shp_file, reference, path_output):

              from osgeo import gdal
              from osgeo import ogr
```

Figure 4. Example of a Jupyter notebook published in EnviDat

For demonstration purposes, the notebook uses fully open vector data for railways, forests and power lines, as well as an open DEM for a small area around a sample forest range in Europe (Germany, Upper Bavaria, Kochel Forest Range, some 70 km south of München, at the edge of the Bavarian Alps). The data of this example can also be used with additional open methodologies and software such as the QGIS plugin Seilaplan (Bont et al. 2018) for optimizing the geometric layout of cable roads or with additional open software for computing the forest accessibility for wood harvesting.

This documented sample dataset allows to demonstrate geospatial preprocessing at the FOSS4G2019 conference based on open data and open software. It contains data needed for computations related to forest accessibility and wood harvesting and was produced based on several existing open data sources.

The vector data (railways, forests and power lines) were extracted from OpenStreetMap (data copyrighted OpenStreetMap contributors and available from https://www.openstreetmap.org) using QGIS; the digitized perimeter of the sample forest range is also made available for reproducibility purposes, although any perimeter or area can be freely digitized (e.g. using the QGIS editing toolbar).

The DEM sample for the selected forest range is based on the open data set collected and resampled by Sonny (sonnyy7@gmail.com) and made available on the Austrian Open Data Portal at http://data.opendataportal.at/dataset/dtm-germany. This original DEM source was reprojected, clipped and resampled to 25 meters using QGIS. More information this example (including data and notebook) is available EnviDat at doi: 10.16904/envidat.75.

5. CONCLUSIONS AND OUTLOOK

Supporting reproducibility of research is a complex issue that can be facilitated by the adoption of FOSS(4G). We would like to stress that reproducibility in science also consists of transparency of methods and accurate documentation of all steps necessary to arrive at the published results. Here, open source software can play a key role.

With EnviDat, WSL fosters international research cooperation in the field of environmental science and contributes to the ongoing cultural evolution in research towards openness, shared data and opportunities for distant collaboration. By openly publishing open software (e.g. as Jupyter notebooks) alongside research data sets, researchers can contribute to mitigate the issues generally described as “reproducibility crisis”. As researchers can combine open source code, detailed text-based descriptions and rich research output in Jupyter notebooks, they will benefit from improved means for opening and documenting their research methods.

In the coming years, EnviDat aims for a deeper integration of user-uploaded Jupyter Notebooks. The development of a generic solution that would offer WSL researchers the opportunity to access their code from anywhere on the existing EnviDat infrastructure is underway. After small-scale proof-of-concept work has been performed that allowed for execution of Jupyter notebooks, we now work on getting such EnviDat-hosted notebooks compatible with the WSL High-Performance Computing (HPC) Linux Cluster. The infrastructure consists of several computing nodes which are connected through a low-latency and high-bandwidth network. Future work will address the issue of providing external access to the current JupyterHub/JuypterLab beta installation on the HPC cluster. Notebooks hosted in EnviDat could then be executed from anywhere, allowing reviewers of WSL research to request access and interactively execute the published code during time windows of low cluster utilization. These developments are expected to strengthen WSL’s commitment to accessible research data in order to advance science and to further evolve EnviDat towards a next-generation environmental data repository. Combined with good and well-established data management practices, these will help provide a better overview over the individual steps of a research process, thus further enhancing the reusability of research results, know-how and expertise.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the support of the WSL IT department, in particular Markus Reinhardt and Stephan Röthlisberger. Furthermore, we would like to thank Matthias Häni and other members of the EnviDat User Group for all their inputs, feedback and support given to the EnviDat program.

REFERENCES

Baker, M., 2016: 1,500 scientists lift the lid on reproducibility. *Nature*, 533, 452–454 (26 May 2016). doi.org/10.1038/533452a

Bont, L. G., Moll, P.E., 2018: Seilaplan. *EnviDat*. doi.org/10.16904/envidat.software.1

Bont, L. G., Fraefel, M., Iosifescu Enescu, I., 2019: Sample Geodata and Software for Demonstrating Geospatial Preprocessing for Forest Accessibility and Wood Harvesting at FOSS4G2019. *EnviDat*. doi.org/10.16904/envidat.75

Fraefel, M., 2018: Dataset for OGRS 2018 publication. *EnviDat*. doi.org/10.16904/envidat.49

Iosifescu Enescu, I., Iosifescu Enescu, C., Pachaud, N. H., Tsorlini, A., Hurni, L., 2015: A decade of geoinformation sharing at ETH Zurich. In Procedings of the *27th International Cartographic Conference: Spatial data infrastructures, standards, open source and open data for geospatial (SDI-Open 2015)*, Rio de Janeiro, Brazil. http://hdl.handle.net/2263/49954

Iosifescu Enescu, I., Plattner, G.-K., Espona Pernas, L., Haas-Artho, D., Bischof, S., Lehning, M. and Steffen, K., 2018(a): The EnviDat Concept for an Institutional Environmental Data Portal. *Data Science Journal*, 17, 28. doi.org/10.5334/dsj-2018-028

Iosifescu Enescu, I., Fraefel, M., Plattner, G.-K., Espona-Pernas, L., Haas-Artho, D., Lehning, M., and Steffen, K., 2018(b): Fostering open science at WSL with the EnviDat Environmental Data Portal. In Proceedings of the *5th Open Source Geospatial Research and Education Symposium (OSGRS 2018), PeerJ Preprints*, Lugano, Switzerland. doi.org/10.7287/peerj.preprints.27211v1

Wilkinson, D.M. et al., 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3:160018 (2016). doi.org/10.1038/sdata.2016.18

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLII-4-W14-107-2019 | © Authors 2019. CC BY 4.0 License.