ENABLING GLOBAL SCALE SENTINEL-1 TIME SERIES ANALYSIS THROUGH STREAMING

Bernhard Raml, Mariette Vreugdenhil, Samuel Massart, Claudio Navacchi, Wolfgang Wagner

TU Wien Department of Geodesy and Geoinformation

ABSTRACT

Dense, high-resolution Synthetic Aperture Radar (SAR) time series from Sentinel-1 offer unique opportunities for monitoring soil moisture. The retrieval process is however challenging due to complex physical processes affecting SAR backscatter, such as vegetation and subsurface scattering effects. Furthermore, the considerable Sentinel-1 data volume introduces its own logistical and computational challenges. This study introduces a novel method for high-throughput calculation of temporal correlation, illustrating how an astute choice of algorithm can facilitate time series analysis on unfavourable data structures. Concretely, we demonstrate how a data streaming approach, with interleaved data reading and processing, can be deployed to efficiently calculate temporal Pearson correlation from a datacube structured as an image stack. Enabled by the substantially reduced computational and memory demands, global calculation of backscatter sensitivity to soil moisture dynamics at a 20 m resolution became feasible. This advancement carries potential for significantly enhancing the accuracy of soil moisture retrievals using SAR backscatter data.

Index Terms— Data assimilation, scalability, high-performance computing, datacube, Sentinel-1

1. INTRODUCTION

Synthetic Aperture Radar (SAR) satellites are a central pillar in Earth observation. Their ability to penetrate clouds, vegetation and soil more effectively than their optical counterparts makes them well suited for continuous all-weather monitoring of the land surface [1]. Moreover, by responding to different physical processes, SAR measurements provide a complementary perspective, revealing features and information that may be obscured or unavailable using other methods [2]. Particularly useful is the Sentinel-1 satellite constellation [3] that has been providing frequent global SAR data coverage at a high spatial resolution of 20 m since 2014, with up to 9 local observations per 12 day repeat cycle depending on the region.

Co-funding by ESA (DTE Hydrology - 4000129870/20/I-NB), FFG (ROSSIHNI - FO999892643, GHG-KIT - FO999893432) and Copernicus (Global Land Monitoring Service - 199494) is acknowledged.

The Sentinel-1 satellites are widely used to retrieve surface soil moisture (SSM) at scales from about 100 m to 1 km [4]. The basis for all current SSM retrieval approaches is that there is a positive relationship between the microwave signal and SSM [5]. However, Wagner et al. [6] showed that this relationship may be inverted in arid and semi-arid environments when the signals penetrate deep into the soil, sensing subsurface scatterers such as rocks and stones. Because of this signal inversion, subsurface scattering leads to negative temporal correlations between the microwave signal and SSM.

Therefore, to identify Sentinel-1 pixels that are potentially affected by subsurface scattering, we compute the temporal correlation between 20 m Sentinel-1 backscatter time series and reference soil moisture data. For our experiment, we chose the ERA5-Land soil moisture dataset, provided by the European Center for Medium-Range Weather Forecasts (ECMWF) at 9 km resolution [7], as a reference. Combining the two datasets at 20 m sampling at a global scale presents the following major challenges: (i) resampling the coarse resolution ERA5-Land data to match the high-resolution Sentinel-1 data efficiently, (ii) handling the sheer volume of the Sentinel-1 data, and calculate the Pearson correlation within a reasonable cost envelope, and (iii) minimizing orbit effects introduced by the Sentinel-1 acquisition method. In the following, we will describe the data, how we resolved these challenges, and finally show first results.

2. DATA

2.1. ERA5-Land

The ERA5-Land dataset is an independently operated land model component of the 5th generation of the European ReAnalysis (ERA5) system, and offers comprehensive global coverage of land surface variables. The data is provided on a 9 km grid with hourly time steps, from which we utilise the SSM, soil surface temperature, and snow depth variables.

Linking the SSM data to the backscatter, requires resampling to match the 20 m grid of the Sentinel-1 datacube. This demands a spatial lookup table, which can consume up to 100 GB RAM, and is computationally intensive. To reduce this overhead, we implemented optimisations in the open-source *pyresample* package[8], described in section 3.1.

doi:10.2760/46796

2.2. Sentinel-1 microwave backscatter

To access high-resolution SAR data, we harness the analysis ready Sentinel-1 σ^0 datacube, as introduced by Wagner et al. [9]. This data is hosted by a dedicated service of the Earth Observation Data Centre for Water Resources Monitoring (EODC, https://www.eodc.eu/), and provided as a stack of compressed GeoTIFF mosaics [9]. The mosaic tiles are given in Equi7Grid projection, introduced by Bauer-Marschallinger et al. [10], and each tile spans an extent of 300km at a resolution of 15000x15000 pixels, resulting in a datacube comprising in total 0.3 petabyte in compressed form.

Handling this large data volume already presents challenges in itself, whereby the spatially-first aligned structure of the datacube makes it particularly difficult to perform time series based analysis. Therefore, caution in the choice of algorithms to calculate temporal parameters, such as Pearson correlation r is advised. To illustrate this, a naive approach would involve decompressing all images contained within an image stack. For a time span of 5 years, this would take approximately 2 hours, and require roughly 12 TB RAM per tile. By implementing a streaming version of the algorithm, we not only significantly reduced RAM usage, but the time spent decompressing can also be completely hidden, as we will show in section 3.2.

To demonstrate our approach we show results calculated over Somalia from 46.7E, 7.6N to 49.6E, 10.4N using Sentinel-1 observations from 2016 to 2021 in section 4.

3. METHODS

Calculating the Pearson r between the σ^0 datacube and ERA5-Land SSM involves the following general processing steps:

1. Resampling and aligning the 9 km resolution ERA5-Land data to match the 20 m resolution σ^0 raster.
2. Masking snow covered and frozen-soil pixels using ERA5-Land *sd*, and *stl1* variables respectively.
3. Calculate Pearson r between σ^0 time series and ERA5-Land SSM based on its *swvl1* parameter for each 20 m pixel per orbit.
4. Compute the average $\bar{r}$ of each orbit's Pearson r weighted by the number of observations.

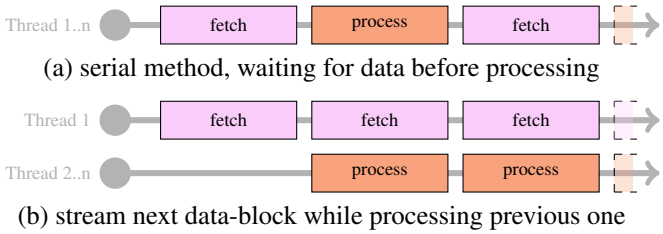
In step 2 we mask out snow covered and frozen-soil pixels because we already know that under these conditions the relationship between $\sigma^0[dB]$ and SSM is undefined. The remaining steps will be described in the following section.

3.1. Data alignment and resampling

To match the ERA5-Land data to the spatial grid of the σ^0 datacube at 20 m resolution, we opted for the open-source Python package *pyresample* to generate the lookup tables. To achieve smooth results, we use Gaussian resampling, considering 8 neighbours with a root mean square width of 9 km, employing the Gaussian resample transformer, published as part of the *eotransform-xarray* [11] package.

The *pyresample* package offers a very convenient API to resample swath data to raster projections, even though the package was not designed with high-resolution SAR images in mind. Computing the lookup table for a raster of this size necessitates a KD-Tree which imposes significant demands on memory. These are difficult to fulfill even for a high performance cluster, hence the lookup table had to be constructed in segments. This revealed a performance issue in *pyresample* related to the concatenation of lookup table segments, but thanks to its very habitable code base, we were able to fix this issue swiftly. This fix is now also publicly available since version *1.27.0*.

For temporal alignment, we use *xarray's* [12] *sel* method to determine the ERA5-Land data timestamp closest to the σ^0 observations.



(a) serial method, waiting for data before processing

(b) stream next data-block while processing previous one

Fig. 1. (a) serial approach where the process sits idle while data is fetched. (b) interleaving data fetch blocks with processing blocks hiding I/O operations.

3.2. Streaming Pearson correlation

As outlined in section 2.2, the σ^0 backscatter data, stored as a stack of GeoTiffs, presents a non-trivial challenge for calculating Pearson r when one has to avoid excessive resource demands. To address the significant RAM requirements, and to utilise the available CPUs more effectively, we have developed the streamed Pearson r algorithm described in Algorithm 1, which is based on the work from Welford [13].

With this streamed approach, we only have to keep the running quantities in memory, independently of the length of the time series under consideration, significantly reducing the memory footprint. Additionally, we can exploit the iterative nature of the algorithm to overlap the I/O communication and decompression with the computation process to efficiently use all the CPUs available. Effectively we use one thread to handle I/O and decompression, while the remaining compute resources of the node can process the data without interruption. This simple streaming strategy pattern has been implemented in the publicly available *eotransform* [14] package, and the

doi:10.2760/46796

Algorithm 1 Streamed Pearson r

$n \leftarrow 0$ ▷ Number of observations
$\mu_a \leftarrow 0$ ▷ Running mean left hand side
$\mu_b \leftarrow 0$ ▷ Running mean right hand side
$M2_a \leftarrow 0$ ▷ Sum of squared differences from lhs mean
$M2_b \leftarrow 0$ ▷ Sum of squared differences from rhs mean
$C \leftarrow 0$ ▷ Co-moment
for each $tile$ in datacube stream **do**
  $n \leftarrow n + 1$
  $\delta_a \leftarrow tile - \mu_a$
  $\delta_b \leftarrow tile - \mu_b$
  $\mu_a \leftarrow \mu_a + \frac{\delta_a}{n}$
  $\mu_b \leftarrow \mu_b + \frac{\delta_b}{n}$
  $\delta_{2a} \leftarrow tile - \mu_a$
  $\delta_{2b} \leftarrow tile - \mu_b$
  $M2_a \leftarrow M2_a + \delta_a * \delta_{2a}$
  $M2_b \leftarrow M2_b + \delta_b * \delta_{2b}$
  $C \leftarrow C + \delta_a * \delta_{2b}$
end for
$r \leftarrow \frac{C}{\sqrt{M2_a * M2_b}}$

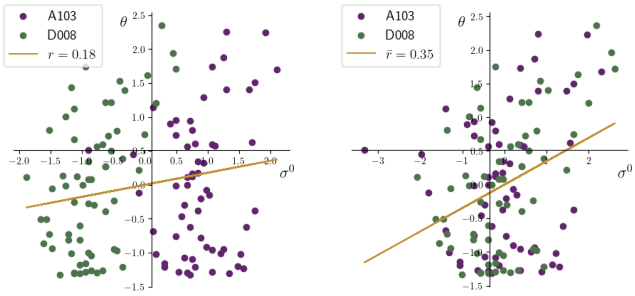
general concept is illustrated in Fig. 1. However, this strategy is only advantageous if loading the data chunks independently is possible, and processing each chunk takes roughly as long as loading it.

Using standard compute nodes of the VSC-4 supercomputer (https://vsc.ac.at/), I/O and decompression of a GeoTIFF takes roughly as long as processing it, resulting in almost no idle time of CPU cores. Consequently, we achieve a throughput of 1.7 GB/s, allowing us to complete a 5-year-long time series of a 15000x15000 pixel tile in 2 hours on a single standard VSC-4 Node, exercising approximately 190 CPU-hours per tile. Furthermore, instead of occupying 12 TB RAM, as the naive approach would, we only require about 7 GB RAM for the correlation process.

3.3. Mitigating orbit effects

In the previous section, we described how we solved the computational problems of processing Pearson r globally. However, one has to consider orbit effects as well, because correlating the full σ^0 datacube for all orbits with the ERA5-Land reference SSM yields lower correlation values than expected. This can be attributed to the fact that observations from different orbits result in distinct incident angles, affecting the measured backscatter.

Fig. 2a illustrates the problem by showing SSM θ and backscatter σ^0 standardised over the full time-series of two orbits A103 and D008 north of Albacete, Spain (-1.53E, 39.47N) from 2020. Fig. 2a demonstrates a clear bias between the two orbits. When interpreting Pearson r as the linear regression slope of these standardised variables, we can clearly see that the correlation is reduced when attempting to



(a) diminished r (b) weighted average $\bar{r}$

Fig. 2. Visualises standardised σ^0 and θ values of two orbits, where (a) demonstrates how differences in incident angles weaken the correlation, while (b) shows the weighted average $\bar{r}$ matching the linear regression slope of separately standardised orbit values.

fit both orbits simultaneously. This low correlation does not accurately reflect the sensitivity of σ^0 to SSM. While individual orbits may exhibit a high sensitivity to SSM, the overall correlation is diminished because incident angles affect the backscatter.

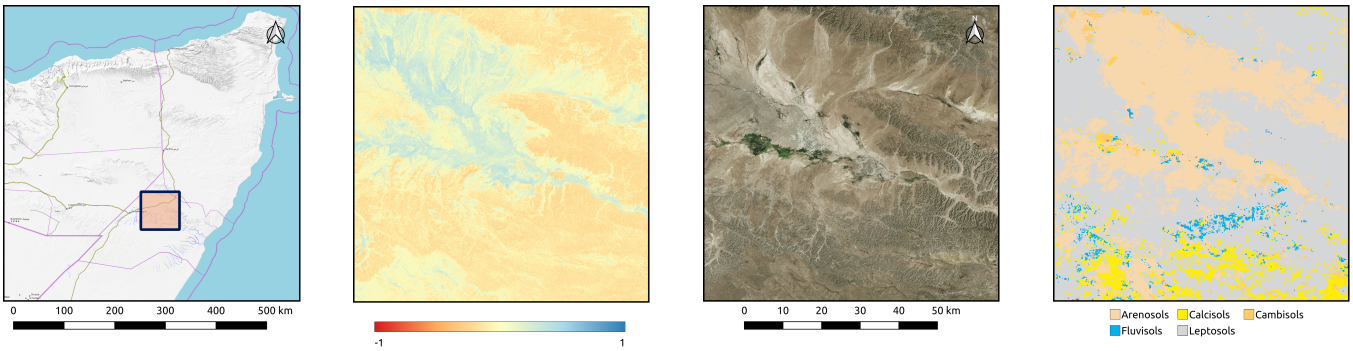
To mitigate the effect, we treat each orbit separately, calculating r_o for each individual orbit. Subsequently, we compute the average $\bar{r}$ weighted by the number of observations per orbit. This weighted average would be equivalent to first standardising σ^0 and θ per orbit and then calculating r. Fig. 2b visualises this, by representing $\bar{r}$ as the linear regression slope of σ^0 and θ independently standardised per orbit. This approach aligns well with our iterative algorithm, as it incurs no additional costs aside from the trivial averaging operation performed at the end.

Another option would be to normalise each orbit to a common reference angle similar to the work of Bauer-Marschallinger et al. [15]. However, this requires a globally accessible slope parameter at 20 m resolution, which is currently unavailable.

4. RESULTS

Fig. 3 shows the resulting Pearson $\bar{r}$ in the region east of Garoowe (48.68E, 8.11N to 49.37E,8.78N), Somalia, in the time from 2017 to 2021. As shown by the accompanying optical image, and soil type maps, this region is characterised by bare or sparsely vegetated ground and soils from the soil groups Arenosols and Leptosols. This presents ideal conditions for subsurface scattering effects to occur, whereas the sign and magnitude of $\bar{r}$ depends on the strength of the subsurface scatterers (rocks, gravel, etc.) and the depth of the intermediate soil layer (shallow soils produce strong negative values). The risk of subsurface scattering effects in the arid regions in the Horn of Africa makes SSM retrieval in this area particularly

doi:10.2760/46796



(a) Region (b) Pearson $\bar{r}$ (c) Bing Aireal (d) SoilGrids soil groups

Fig. 3. (a) Location of examined region in Somalia, (b) Pearson $\bar{r}$ map east of Garoowe, (c) optical image, and (d) soil groups.

challenging. With the capability of identifying pixels insensitive to SSM, we aim to improve existing medium-resolution (1 km) SSM products over Africa and similar regions.

5. OUTLOOK

We are now preparing the production of a complete global dataset, to be analysed and validated in further studies. Furthermore, this particular algorithm can be extended to generate correlation maps capturing seasonal SSM sensitivity.

Currently, we are using standard compute nodes from the VSC-4 supercomputer, yet, this method is particularly well suited to high throughput optimised GPU hardware. This may further reduce costs and energy consumption, and could be the subject of future work.

REFERENCES

[1] F.T. Ulaby, R.K. Moore, and A.K. Fung. *Microwave remote sensing: Active and passive. Volume 1-microwave remote sensing fundamentals and radiometry*. Artech House microwave library. Artech House Inc, 1981.

[2] M. Vreugdenhil, W. Wagner, B. Bauer-Marschallinger, et al. Sensitivity of sentinel-1 backscatter to vegetation dynamics: An austrian case study. *Remote Sensing*, 10(9), 2018. doi: 10.3390/rs10091396.

[3] R. Torres, P. Snoeij, D. Geudtner, et al. Gmes sentinel-1 mission. *Remote Sensing of Environment*, 120:9–24, 2012. doi: 10.1016/j.rse.2011.05.028. The Sentinel Missions - New Opportunities for Science.

[4] B. Bauer-Marschallinger, V. Freeman, S. Cao, et al. Toward global soil moisture monitoring with sentinel-1: Harnessing assets and overcoming obstacles. *IEEE Transactions on Geoscience and Remote Sensing*, 57(1):520–539, 2019. doi: 10.1109/TGRS.2018.2858004.

[5] M.C. Dobson and F.T. Ulaby. Active microwave soil moisture research. *IEEE Transactions on Geoscience and Remote Sensing*, GE-24(1):23–36, 1986. doi: 10.1109/TGRS.1986.289585.

[6] W. Wagner, R. Lindorfer, T. Melzer, et al. Widespread occurrence of anomalous c-band backscatter signals in arid environments caused by subsurface scattering. *Remote Sensing of Environment*, 276:113025, 2022. doi: 10.1016/j.rse.2022.113025.

[7] J. Muñoz Sabater, E. Dutra, A. Agustí-Panareda, et al. Era5-land: a state-of-the-art global reanalysis dataset for land applications. *Earth System Science Data*, 13(9):4349–4383, 2021. doi: 10.5194/essd-13-4349-2021.

[8] David Hoese, Martin Raspaud, Panu Lahtinen, et al. pytroll/pyresample: Version 1.27.0 (2023/05/17). May 2023. doi: 10.5281/zenodo.7943813.

[9] W. Wagner, B. Bauer-Marschallinger, C. Navacchi, et al. A sentinel-1 backscatter datacube for global land monitoring applications. *Remote Sensing*, 13(22), 2021.

[10] B. Bauer-Marschallinger, D. Sabel, and W. Wagner. Optimisation of global grids for high-resolution remote sensing data. *Computers & Geosciences*, 72:84–93, 2014. doi: https://doi.org/10.1016/j.cageo.2014.07.005.

[11] Bernhard Raml. eotransform-xarray. June 2023. doi: 10.5281/zenodo.8002854.

[12] Stephan Hoyer, Maximilian Roos, Hamman Joseph, et al. xarray. May 2023. doi: 10.5281/zenodo.7949546.

[13] B. P. Welford. Note on a method for calculating corrected sums of squares and products. *Technometrics*, 4(3):419–420, 1962. doi: 10.1080/00401706.1962.10490022.

[14] Bernhard Raml. eotransform. June 2023. doi: 10.5281/zenodo.8002789.

[15] B. Bauer-Marschallinger, S. Cao, C. Navacchi, et al. The normalised sentinel-1 global backscatter model, mapping earth's land surface with c-band microwaves. *Scientific Data*, 8(1):277, 2021.