Sentinel-1 Data Preprocessing

The incoming Level-1 Sentinel-1 datasets must undergo the preprocessing routines before forwarded to the algorithms for water and flood extent detection. The output of the preprocessing is Analysis-Ready-Data (ARD) that is formatted and gridded, and immediately forwarded to GFM flood detection engine and added to the Sentinel-1 data cube. The preprocessing of Sentinel-1 data constitutes the by far most computation effort in the entire processing chain, only achievable in NRT when employing a large-scaled computing cluster, efficient job parallelization, a fast file system, and a capacious data storage. In detail, we ingest into the preprocessing module the observations from the Sentinel-1A/B satellites that are acquired in Interferometric Wide-swath mode and Ground Range Detected at High resolution (Sentinel-1 IW GRDH). The GRD products consist of focused SAR data that has been detected, multi-looked and projected to ground range using an Earth ellipsoid model, and phase information is lost. The resulting product has approximately square spatial resolution pixels and square pixel spacing with reduced speckle at the cost of worse spatial resolution. In case of the here used high-resolution product, the raw backscatter amplitude is sampled with a 10x10 m pixel size. For the GFM flood service, we process Sentinel-1 data in VV-polarization (and neglect the VH-polarization channel) due to its higher sensitivity in differentiating water from non-water surfaces.

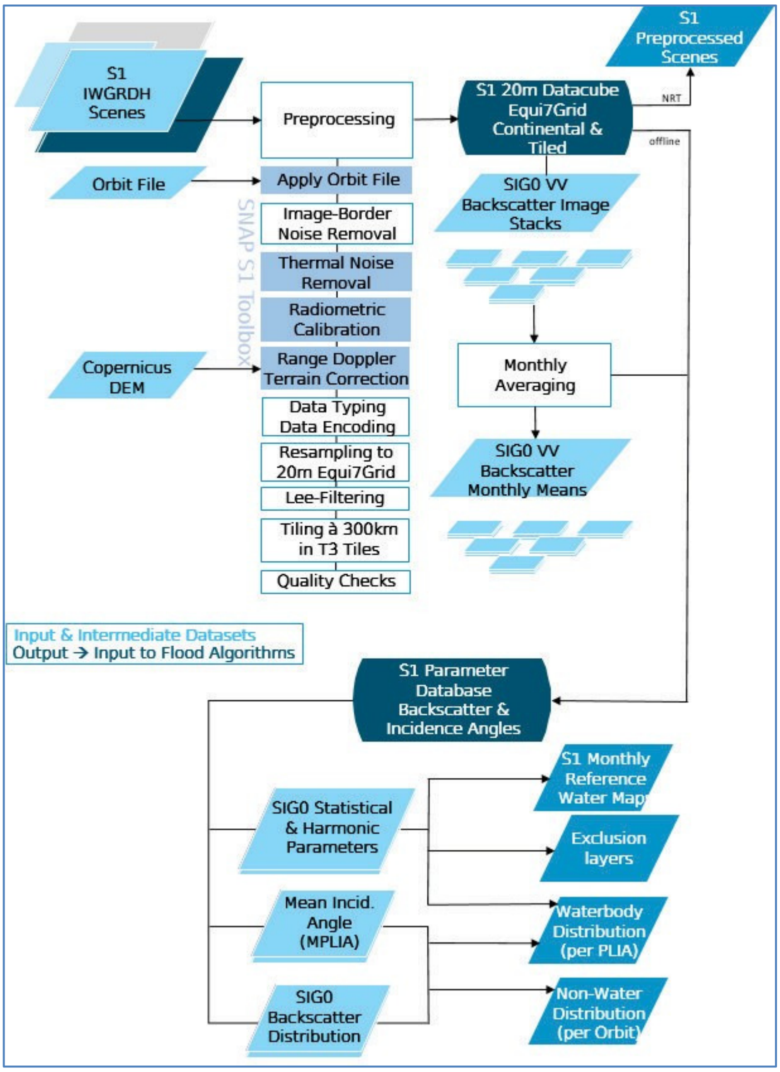
Most beneficial, a NRT operational preprocessing chain has been already implemented at Vienna by EODC together with TU Wien, facilitating the Sentinel-1 input dataflow for the soil moisture products in the Copernicus Global Land Service (CGLS). In the CGLS, the 1km Surface Soil Moisture from Sentinel-1 (SSM1km) and the 1km Scatterometer-SAR-Soil Water Index (SWI1km) are generated over Europe since 2019 and are currently in preparation for rolling-out to global coverage. As both CGLS soil moisture products and the here proposed GFM product are based on preprocessed Sentinel-1 IW GRDH observations in VV-polarization, we can exploit knowledge,experience, and hard- and software infrastructure in a most synergistic and efficient manner.

Regarding SAR processing software, the GEO Department of TU Wien developed the dedicated software suite SAR Geophysical Retrieval Toolbox (SGRT v2.4) for the large-scale processing of Sentinel-1 products and the subsequent data aggregation and extraction of geophysical parameters. Its framework is in Python language and comprises workflows on several image and signal processing components, including a parallelized SAR preprocessing module environing ESA's state-of-the-art engine Sentinel Application Platform (SNAP v6.0). The TU Wien preprocessing module ingests the Sentinel-1 IW GRDH images, and applies for each scene the following operators of the SNAP software:

- orbit vector correction, using “POD Restituted Orbit Ephemerides” provided by ESA.
- thermal noise removal
- radiometric calibration, using the information from the Sentinel- 1Product Auxiliary File
- range Doppler geometric terrain correction using the Copernicus Digital Elevation Model (CopDEM)

For the terrain correction, the newly available Copernicus DEM will be used, as requested by the call. From the technical specifications, it can be considered as benchmark for global surface and terrain modelling and appears most suitable for the processing of Sentinel-1 in the GFM service. The CopDEM is available for the global extent in two spatial resolution qualities (90m and 30m, with the latter available for eligible entities and usages). As the CopDEM is new to the community, it must be evaluated if it is fully suitable for range Doppler geometric terrain correction in SNAP, recognizing that voids and artefacts can have severe impact on the preprocessing output. Also, the data type of the CopDEM must be closely investigated, as it is a digital surface model (DSM) according to the documentation. Detailed information on the surface height potentially enhances the accuracy of the terrain correction in the SAR perspective and could further improve the modelling of radar shadows along forest lines and in the vicinity of high buildings. As a reliable fallback solution, the currently 90m fully global DEM used at the TU Wien / EODC preprocessing engine can be employed in the GFM service, which is based on the 3-arc seconds SRTM, the GDEM, and topographic maps (cite Ferranti). It features no serious voids nor artefacts and has proven compatibility in the Sentinel-1 NRT operations of the CGLS.

As additional preprocessing step, a statistically based image-border-noise-removal is applied, following an algorithm developed specifically for Sentinel-1 (Ali2017). The Sentinel 1 image-border noise affects the Level-1 GRD products older than 2017, as this issue was resolved by the Sentinel 1 CSAR Instrument Processing Facility (IPF) version 2.9. The artefacts at the image borders are especially harmful to statistical backscatter parameters, with corrupt pixels values that contaminate the time series at an irregular pattern. The operator provided by SNAP for removing image-border noise is unfortunately not consistent and fails to remove the artefacts completely. The here included noise removal procedure is a “bi-directional” noise removal method that removes the border noise completely without losing adjacent valid pixels.


S1 IWGRDH Scenes
Preprocessing
S1 20m Datacube Equi7Grid Continental & Tiled
S1 Preprocessed Scenes
NRT
offline
Orbit File
Apply Orbit File
Image-Border Noise Removal
Thermal Noise Removal
Radiometric Calibration
SNAP S1 Toolbox
Copernicus DEM
Range Doppler Terrain Correction
Data Typing Data Encoding
Resampling to 20m Equi7Grid
Lee-Filtering
Tiling à 300km in T3 Tiles
Quality Checks
SIG0 VV Backscatter Image Stacks
Monthly Averaging
SIG0 VV Backscatter Monthly Means
Input & Intermediate Datasets
Output → Input to Flood Algorithms
S1 Parameter Database Backscatter & Incidence Angles
SIG0 Statistical & Harmonic Parameters
Mean Incid. Angle (MPLIA)
SIG0 Backscatter Distribution
S1 Monthly Reference Water Map
Exclusion layers
Waterbody Distribution (per PLIA)
Non-Water Distribution (per Orbit)

General workflow for building the Sentinel-1 datacube, comprising the 20m Level-1 sig0 backscatter observations(Analysis-Ready-Data, ARD), and the therefrom derived statistical and temporal parameters used as input to the flood algorithms.

After application of the SNAP operators, the SAR data is ingested into the Sentinel-1 data cube. The ingestion involves (with details below):

- Resampling to the projection of the Equi7Grid, at 20m spatial sampling
- Adaptive Lee-Filtering
- Data type conversion step
- Geographic tiling into 300km² boxes ("T3-Tiles") defined by the tiling scheme of the Equi7Grid
- Automated checks on SAR output quality and file health

The intermediate geocoded Sentinel-1 scene output from SNAP is given in the native latitude-longitude projection, for performance and consistency reasons. Using the GDAL libraries, the data is resampled to the Equi7Grid projection system with a spatial sampling of 20m pixel size. For the geographic image-warping the bilinear resampling method is used. The resampling from an initial 10m sampling to a 20m sampling is a measure with great impact on the GFM service and the delivered product. The reasoning for this downsampling is because a Sentinel-1 GRDH image, owing to the nature of the SAR observation technique, inevitably carries speckle and signal noise. The latter make that the effective resolution is somewhat coarser than the nominal resolution. A Sentinel-1 image with 20 m square pixels has four equivalent number of looks (ENL) and, as a result, the speckle standard deviation is halved and the PDF becomes Gaussian (Xie et al., 2002). Many automatic floodwater mapping algorithms take advantage of both effects. Indeed, the information content on surface characteristics in a ground range detected SAR image is depending on the local speckle/noise level and the complexity of the observed surface structure. Hence, the resampling does not impact the accuracy of the generated water and flood extent maps. The visualization examples in the Annex 5 and the comparison between 10 m and 20 m products in Figure 25 illustrate the appropriateness for flood mapping based on 20m-sampled Sentinel-1 backscatter. Moreover, and most importantly for the NRT operations of the GFM service, the reduction of data volume resulting from the downsampling is substantial, with a reduction factor of approximately between 3 and 4 in total required storage. Hence, the reduced data volume helps significantly reducing the cost for purchasing and maintaining the operator's storage capacities, and further improving the timeliness of the flood product. With less data (i.e., arrays of smaller size) to be forwarded and processed by the flood detection algorithms, the computation time and all downstream operations to final production is speeded up accordingly.

Contrast and thus helps to accentuate individual features. This is most useful for the mapping of water surfaces, with their strong contrast in backscatter to non-water surfaces. Ideally, the Lee-filter is applied to the initial 10m-sampled data, but an analysis carried out by TU Wien has shown that the difference between the sequence filtering-resampling and the reversed sequence (resampling-filtering) is minimal in terms of resulting spatial backscatter signal but differ in computation time by factor 2-3.

The data conversion involves 16bit-encoding of the backscatter values expressed in Decibels, a linear value scaling to fully exploiting the available integer value range, and a file compression using the Z-standard (zstd) method. The latter is an open-source lossless data compression algorithm based on indexing reoccurring sequences in data, using latest entropy coding techniques aimed for high performance. Especially for data with a limited value range such as the here used Sentinel-1 IW GRDH SAR data, significantly better compression rates can be achieved compared to similar compression methods. Homogenous image regions in SAR data, in particular desserts or water bodies allow efficient compression of large sequences. In terms of compression performance, a 20% advantage of zstd against the previous state-of-the art compression LZW has been achieved within the Sentinel-1 ARD backscatter data cube. The open zstd package also includes parallel (multi-threaded) implementations of both compression and decompression, enabling faster file I/O operations compared to LZW. With this, we can assure a most efficient way forward in respect to necessary storage volume and I/O access.

In the next step, the resampled and filtered backscatter images are tiled into “T3”-tiles of the Equi7Grid (squares covering 300x300km). This is done to keep balance between the file- and array- sizes, which must not be too large to be manageable during runtime, and the number of individual files, which should keep as low as possible to minimize overhead in the operating file system. The tiling also supports time series access via fast pixel indexing and multitemporal image stacking.

After the tiling, the data from an incoming Sentinel-1 scene undergo a quality and consistency check, flagging abnormal outputs that are disregarded in downstream processing. Finally, they are forwarded to the data cube and the flood detection algorithms for NRT production.

With this, the Sentinel-1 data cube is constantly updated, but it also comprises SAR observations starting with Sentinel-1 mission launch in 2014. In an offline-processing framework, these datasets are used for the generation of monthly mosaics and statistical parameters.