ESA CONTRACT REPORT

Contract Report to the European Space Agency

SMOS Near-Real-Time Soil Moisture processor: Operational chain and evaluation

J. Muñoz-Sabater, N.J. Rodríguez-Fernández, P. Richaume, C. Albergel, P. de Rosnay, Y.H. Kerr

ESA/ESRIN contract 4000101703/10/NL/FF/fk.

Technical Report 2 (TR2) Work Package 4020

European Centre for Medium-Range Weather Forecasts
Europäisches Zentrum für mittelfristige Wettervorhersage
Centre européen pour les prévisions météorologiques à moyen terme

ECMWF

Series: ECMWF ESA Project Report Series

A full list of ECMWF Publications can be found on our web site under:
http://www.ecmwf.int/en/research/publications

Contact: library@ecmwf.int

European Centre for Medium Range Weather Forecasts
Shinfield Park, Reading, RG2 9AX, England

Contract Report to the European Space Agency

SMOS Near-Real-Time Soil Moisture processor: Operational chain and evaluation

Authors: J. Muñoz-Sabater, N.J. Rodríguez-Fernández, P. Richaume[1], C. Albergel, P. de Rosnay, Y.H. Kerr[1]

ESA/ESRIN contract 4000101703/10/NL/FF/fk.
Technical Report 2 (TR2) Work Package 4020

[1] CESBIO (CNRS, CNES, IRD, Université Paul Sabatier), 18 av. Edouard Belin, bpi 2801, 31401 Toulouse cedex 9, France

January 2016

	Name	Company
First version prepared by (January 2016)		ECMWF
Quality Visa		ECMWF
Application Authorized by		ESA/ESRIN

Distribution list:

ESA/ESRIN
Susanne Mecklenburg
ESA ESRIN Documentation Desk

SERCO
Raffaele Crapolicchio

ESA/ESTEC
Matthias Drusch

ECMWF
HR
Division & Section Heads

Contents

1 Introduction

This document describes the technical implementation of a new soil moisture (SM) product in Near-Real-Time (NRT), based on SMOS NRT brightness temperatures (TBs) and the training of a neural network (NN) using SMOS Level 2 SM as reference. Hereupon this product will be named as SM-NRT-NN. The general methodology has been established during the SMOS+"Neural Network" project [13]. The selection of the the best NN architecture, the algorithm specifications and the evaluation with respect to SMOS Level 3 CATDS[1] data can be found in the report [14]. In addition to the operational implementation, the current report also discusses the evaluation of the SM-NRT-NN product. The implementation of the SM-NRT-NN processor has been conducted as follows:

1. Implementation of a TB processor and generation of a large database of angle-binned TBs in horizontal and vertical polarization from NRT TBs. The full TB angular profiles are sequentially created per SMOS Discrete Global Grid (DGG) and for a pre-defined training period.

2. Retrieval from the ECMWF MARS archive of forecasted soil temperature information at the time of the observation and interpolation to the SMOS DGG grid for each orbit.

3. Collocation of the angle-binned NRT TBs, forecasted soil temperature and snow depth with SMOS Level 2 SM and retraining of the NN. Computation of NN parameters.

4. Implementation of the NN parameters and operational production.

5. Evaluation of the SM-NRT-NN product.

A detail flow diagram of the whole processor has been presented in [14]. The operational phase is conducted in NRT, by half-hourly checking for the arrival of new orbits deposited in the ECMWF raw satellite data repository. Each time a new orbit is found, the operational chain is triggered.

Section 2 describes each technical step in the production of the SM-NRT-NN product and the setup of the operational chain. Instead of using the NN weights obtained with the Level 3 prototype and discussed in [14], to ensure the optimal performances for the SM-NRT-NN product, the NN has been retrained with actual NRT TBs and using SMOS Level 2 SM (version 6.20)as reference data. The training of the NN is discussed in Section 3. The SM-NRT-NN product has been evaluated against SMOS Level 2 SM in swath-basis and globally during two periods: (*i*) from June 2010 to June 2012 using reprocessed data (*ii*) from May to November 2015 using actual operational TBs. This evaluation, in addition to comparison with in situ measurements in North America, is discussed in Sect. 4.

2 Implementation

2.1 Processor structure and tasks

The full processor of the SM-NRT-NN product is based on the sequential execution of several shell scripts and Fortran routines as shown in Fig. 1. The main tasks of each component (or subprocessor) of the full processor are listed below:

[1] *Centre Aval de Traitement de Données SMOS*

1. Master processor (master.sh): This is the top-level script and it runs in the ECMWF linux cluster. It is triggered once every half an hour and controls the communication with the linux-cluster (maximum processing time, notification emails, etc.). It does all the time management of old and new orbits, prepare the setup of the processor, check for delayed files (too old files will not be processed), clean directories with old data, etc. It triggers the production phase if a new orbit in the ECMWF SMOS data repository is found. This script allows new orbits to be processed if the NRT files arrive up to a maximum delay of five hours after midnight UTC. The master is executed with a frequency which is customizable, as well as the parameter controling the maximum orbit time delay.

2. Operational processor (oper_ecmwf_processor.sh): This script is called by the master and it controls the main processing of the data. The main tasks performed are the following:

   - To retrieve SMOS NRT data from the ECMWF ECFS repository, which is the repository where the raw SMOS NRT BUFR data is pushed and archived. For the training phase of this document, the second reprocessed dataset of the SMOS mission is used, covering the year 2010 to 2014. The reprocessed data is archived in ECFS (note: at the time of writing this report, the forecast department plan to retrieve and archive these data too). By defining the variable ”$DATA”, the user of the processor defines where the script looks for the data. It is straightforward to implement alternative archives where to find and retrieve the NRT files.
   - The Fortran compilation is setup here (flags, processors, etc.). Then, all the Fortran source is compiled and the executables produced,
   - To prepare time and step of auxiliary fields forecasts (soil temperature, soil depth),
   - To trigger the processor of the auxiliary fields,
   - To test the retrieved forecast fields based on the date of the retrieval,
   - To trigger the final production of the SM-NRT-NN product in netcdf format,
   - To rename the final product using the convention name,
   - To push the SM-NRT-NN product to ESA ftp if desired.

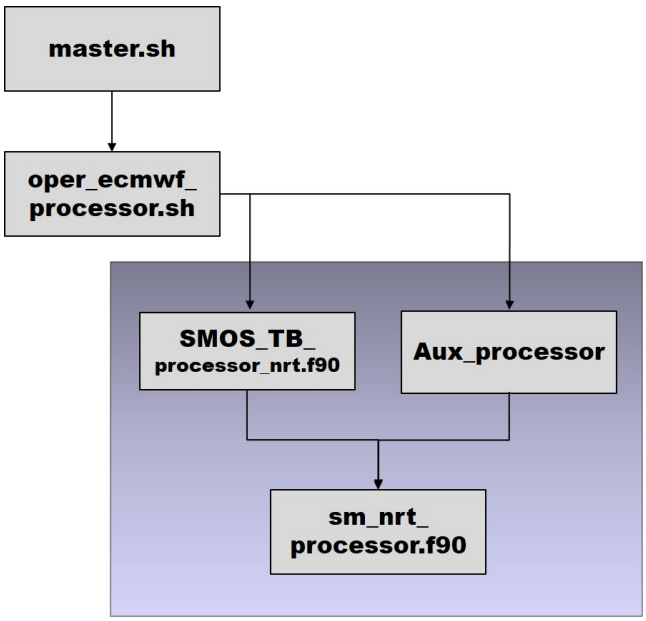

Figure 1: Flow diagram of the production of the SM-NRT-NN product

3. SMOS brightness temperatures processor (SMOS_TB_processor_nrt.f90): This is a Fortran-90 routine, and it is the longest component of the full processor. This routine is responsible for performing the transformation of an orbit from NRT brightness temperatures to CATDS L3TB-like [12, 5] brightness temperatures. The main tasks associated to this processor are the following:

   - All the snapshots of a single orbit are distributed between a number of s-processors. The distribution is done by the master processor (m-processor)
   - Each s-processor processes sequentially a list of snapshots and does the following subtasks:
     (a) Decodes 'raw' BUFR snapshots (this is a big advantage, as there is no need to wait for the availability of 6 h pre-processed BUFR files produced by the ECMWF Forecast Department, what would increase substantially the latency of the SM-NRT-NN product),
     (b) Assigns land DGG numbers and auxiliary data needed for further processing,
     (c) Computes time information with regards to a reference (01 January 2000),
     (d) Filters observations (Radio Frequency Intereference (RFI), sun-tails, suspicious values, ocean, etc.). The different filters can be switched on/off by a logical variable,
     (e) Collects all observations for each selected DGG,
     (f) Constructs brightness temperatures profiles (s-tables) with a subsample of snapshots,
   - The m-processor collects all s-tables and construct an unique table (m-table) with all the brightness temperatures profiles per DGG for further processing,
   - Evaluate the presence of breaks in a sequence of snapshots based on the polarisation value and the time difference between a sequence of observations (break-hit table),
   - For each segment of consecutive (and sufficient number of) snapshots, the full profile of brightness temperatures and radiometric accuracy in XX and YY mode is computed using linear interpolation (extrapolation is not permitted),
   - Then, the rotation to the Earth reference frame ([X,Y] to [H,V]) using the full MR4 matrix is conducted following [12] as recommended by [14].
   - Data is binned per DGG grid point and averages are computed per angular bin,
   - Further quality control is applied to single profiles,
   - Output is written out for further processing,
   - Statistics for this orbit are computed and displayed on the screen.

   It also should be noted that at the time of this report, the paralelisation of the snapshots processing is equivalent to a sequential paralelisation. The reason is that the interpolation is not paralelizable per chunks of random snapshots. With the current optimization, a sequential processing is still very performing.

4. Auxiliary fields processor (aux_processor): This scripts is coded in metview macro code and it is embedded within the operational processor. It generates auxiliary fields collocated to the SMOS DGG grid (only grid points in the orbit). The main tasks carried out by this script are:

   - Retrieval of the following fields from the MARS archive, with parameters set by the operational processor:
     (a) Soil temperature,
     (b) Snow depth,
     (c) Land surface mask.
   - It is based on the closest operational forecast available in time to the observations. If the closest forecast is not available, then it searchs for the previous one,

- Collocates the auxiliary data to the SMOS DGG grid points of the orbit being processed. The nearest neighbour approach is adopted, as the ECMWF high-resolution forecast fields (approximately 16 km) and the SMOS land DGG grid are very close in spatial resolution at the time of this report,
- Produces a netcdf auxiliary field, used by the SM NRT processor.

5. SM NRT processor (sm_nrt_processor.f90): This is the final routine, coded in Fortran-90. It generates the final soil moisture product with uncertainties and probability of Radio-Frequency-Interference fields. The main tasks conducted by this routine are the following:

   - Set up the neural network coefficients.
   - Read and store information needed to process the final product (profiles, auxiliary fields, max/min tables, etc.).
   - Computes the soil moisture product based on the Neural Network weights and auxiliary fields for each DGG of the orbit being processed.
   - Computes quadratic errors of the NN output based input uncertainties.
   - Creates an output netcdf with all the required information.
   - Check the final netcdf.

   More information on the uncertainty computation and the algorithms behind the soil moisture product can be found in [14].

2.2 Technical characteristics of the full SM-NRT-NN processor

The technical details of each component of the full processor are given in this section:

- master.sh: 160 lines of shell script code,
- operational processor (including aux_processor): 520 lines of shell script code (last version is v1.3)
- SMOS brightness temperatures processor: 2250 lines of Fortran-90 code (the version described here is 1.15), and composed of 4 subroutines and 5 functions.
- SM NRT processor: approximately 1000 lines of Fortran code (last version is 1.11).

All in all, the full processor is worth 3250 lines of Fortran code and approximately 700 lines of shell script. The full product is processed in approximately 1 minute, but this figure varies depending on the size of the orbit.

2.3 Description of the SM-NRT-NN product

The SM-NRT-NN product is an only land product, colocated and delivered in the SMOS DGG grid, i.e., the ISEA 4H9 grid [15]. The format selected is NetCDF (version 4.0). The main characteristics of the product and the description of the fields are presented in Tables 1 and 2.

Filters

Each TB profile is transformed into a SM value and disseminated if it survives to the following filters applied to the individual TBs before averaging, to the angle-binned TB versus incidence angle profiles or to the output SM themselves:

1. **Hard RFI**: Each observation which is not within the observed physical range ($80K < TB(K) < 340K$) is considered to be corrupted or affected by RFI. In this case the observation is not used to produce a soil moisture estimate.
2. **Cross-polarised values out of the expected range**: The real and imaginary components of the cross-polarised observations are also checked for suspicious values. If any component is not within these ranges $[-50K < Re(TBXY) < 50K]$ and $[-50K < Im(TBXY) < 50K]$, then the observation is not used to produce a soil moisture estimate.
3. **Sun-alias reconstruction**: The observed TB is filtered out if a specific BUFR flag indicates that the observation is located in a zone where a Sun alias was reconstructed.
4. **Number of observations per angular bin**: a soil moisture estimate is not produced if at least one observation (brightness temperature) is not available in each angular bin.
5. **Snow cover**: soil moisture is filtered out if snow is found in the grid point, which is based on the latest ECMWF snow depth forecast field. For most of the files, the time difference between the soil moisture estimate and the snow depth forecast is below 1 hour.
6. **Freezing soil**: soil moisture is filtered out if the soil is frozen. To this end, if the soil temperature forecast of the top soil 7 cm is below 274 K, then a soil moisture estimate is filtered out. For most of the files, the time difference between the soil moisture estimate and the soil temperature forecast is below 1 hour.
7. **Land-surface mask**: soil moisture is filtered out if more than 50% of a pixel is covered by water. This filter avoids too 'wet' values near the coastlines.

It is also important to notice that some swath-files of the final product will be empty (total size of 796 K). This is common in the case that the Near-Real-Time file processed by the full processor has a very low size, with a few number of observations, and none of them have survived to the previous filters and quality control steps. In the case all the values are filtered out due to snow or frozen soil, then the final product will still be delivered for information, but no soil moisture estimate will be available. Obviously, this behaviour will be more common in the winter of the North Hemisphere than in summer, as more grid points are affected by snow or frozen soil.

Product Description	Product name	SMOS Near-Real-Time soil moisture
	Acronym	SM-NRT-NN
	Product Description	SM-NRT-NN product provides a Near Real Time soil moisture dataset based on the statistical weights computed by a Neural Network. It is provided in the SMOS DGG grid and only at the satellite track. It also provides an estimation of the soil moisture product uncertainty and the probability that a soil moisture value is contaminated by Radio Frequency Interference. It is distributed in NetCDF format and it is intended to be used for a wide range of applications requiring NRT information on soil moisture.
Geographical Bounding Box (the product is delivered in the satellite track, which cover the following geographical bounds)	West bound longitude	-180
	East bound longitude	180
	North bound latitude	90
	South bound latitude	-90

Table 1: General features of the SM-NRT-NN

NetCDF Field	Units	Description
Latitude	degrees	Geographic latitude of the retrieval
Longitude	degrees	Geographic longitude of the retrieval
Soil moisture	m^3m^{-3}	Soil moisture retrieval value
Soil moisture error	m^3m^{-3}	Estimated uncertainty of the retrieval
RFI probability	%	Probability that the retrieval is affected by RFI. The RFI probability is computed per DGG grid point, and it is defined as the number of BUFR brightness temperature observations flagged as affected by RFI with respect to the total number of observations. Only those observations remaining after filters 1, 2 and 3 of section 2.3 are accounted for in the RFI probability calculation. In the BUFR NRT TB product, three bits of the SMOS information flag indicate the following information about RFI: bit-1:*"Pixel is affected by RFI effects as identified in the AUX_RFILST or it has exceeded the BT thresholds"*", bit-4:*"Measurement is affected by the tails of a point source RFI as identified in the AUX RFI list (tail width is dependant on the RFI expected BT, from each snapshot measurements, corresponding to 0.16 of the radius of the RFI circle flagged)"*, and bit-9:*Measurement is affected by RFI effects in the corresponding polarisation as identified in the long trend analysis of telemetry data (NIR and System Temperatures).*"
Number of days since 01-01-2000	days	Number of days since 01 January 2001
Number of seconds since midnight UTC	seconds	Number of seconds since 00UTC
SMOS DGG id	-	Identification number in the SMOS DGG grid

Table 2: Description of the fields included in the SM-NRT-NN product

2.4 Running the processor

The NRT processor can be run as simply as executing the top level master script. There are a series of hard-coded parameters inside the scripts and routines, and which can be modified to control different aspects of the processor, as for instance the angular bins, the maximum number of observations allowed per DGG, etc. The processor has also been prepared to be run in the linux cluster, for which a header in the master script is needed to interface with the linux cluster. Here below there is a series of rules that have to be checked in order to make the processor portable to other workstations:

1. Set paths:
   - master.sh
     - Set **$mydir** → location where source scripts and Fortran routines are found
     - Set **$outdir** → location where all orbit processed are written to check double-processing
   - operational processor
     - Set **$mydir** → location where source scripts and Fortran routines are found
     - Set **$SOURCEf90** → directory where all Fortran code is accessed
     - Set **$SOURCEgptHEADER** → directory where header for metview geopoints is found
     - Set **$SOURCEmetview** → directory where metview source is found
     - Set **$WDIR** → working directory
   - SMOS brightness temperatures processor
     - Set full path of 'isea4h9_land.txt' and 'FWOLand.txt' aux files
   - SM NRT processor
     - Set full path for auxiliary files containing the NN coefficient.
     - Set full path for the '**FILENAME**' variable, in unit 8,
     - Set full path for '**isea4h9_land.txt**' file
     - Set full path for the '**COORDIDGGNATE'** variable, in unit *iunitcoord*
2. Set check variables for debugging to false:
   - SM NRT processor
     - Set **LCHECK**=.FALSE.
3. ftp settings: the server, user and passwords have to be set if dissemination to an external ftp is needed;
   - operational processor
     - Set the variable **lpushstp**=true
     - Set ftp credentials, either setting the variables **$HOST**, **$USER** and **$PASSWD**, or using a proxy machine.
4. Set crontab to run master script automatically every some specific time. Currently it is set up to be run every 30 minutes
5. Check all module loads, as this is important to look for the up to date libraries
6. Check the Fortran compilation, especially the compiler used.

Other modifications maybe necessary in order to integrate this software in other architectures, for example using SMS suites.

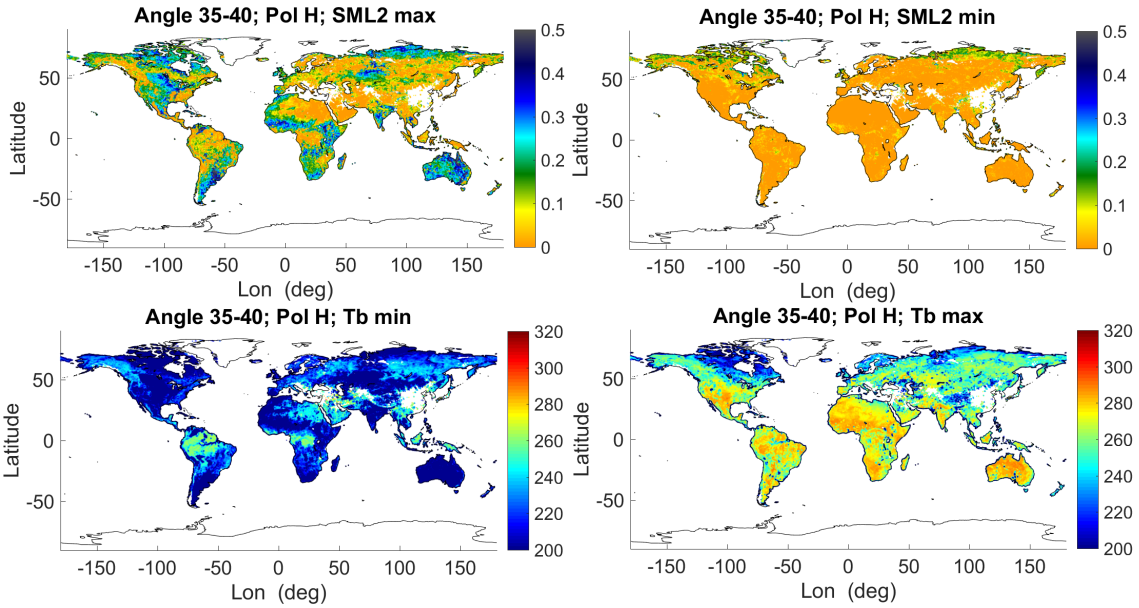

Figure 2: Example of extreme TBs values for H polarization and angle bin 35°-40° and associated SMOS L2 SM.

3 Retraining the Neural Network

Using reprocessed NRT TBs in BUFR format [7, 6] from 01/06/2010 to 30/06/2012, the SM-NRT-NN processor has been used to compute a training data base with TBs in H and V polarizations and three incidence angle bins from 30° to 45°. As described in [14], this configuration is the best trade-off of retrieval accuracy and swath-width. The NRT-NN SM is retrieved in swaths of $\sim$ 915 km (the standard L2 algorithm [10] retrieves SM in swaths of $\sim$ 1150 km).

The HV angle-binned TBs have been collocated with ECMWF IFS forecasts for the soil temperature and snow cover and finally they have been collocated with version 6.20 SMOS L2 SM data [10, 9]. Following the algorithm description in [14] a local normalized index (hereafter I_2) has been computed from extreme TBs and associated L2 SM for each DGG point. To compute the extreme values tables, the following filters have been used:

- The latitude is limited to the [-60°,75°] range.
- Water fraction within the SMOS footprint is required to be zero.
- A SMOS L2 SM value associated to the maximum or minimum TB is required (otherwise I_2 cannot be defined).
- The SM uncertainty given by `Dqx` parameter should be lower than 0.06 m^3/m^3.
- The soil temperature should be higher than 274 K to avoid frozen soil (taking into account a temperature uncertainty of $\sim$ 1 K.
- The snow depth should be zero.

Figure 2 shows an example of the extreme values tables for H polarization and angle bin 35°-40°. Figure 3 shows the associated uncertainties.

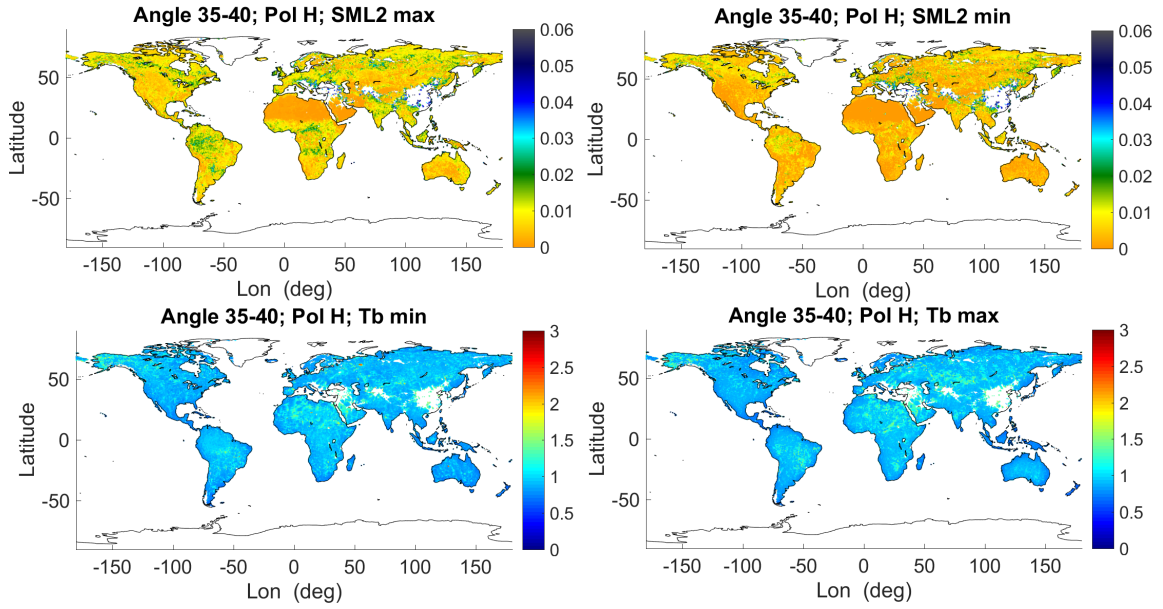

Figure 3: Uncertainties associated to the maps shown in Fig. 2.

Following the study during the first phase of this project using the NRT prototype [using SMOS CATDS L3 data, see 5]. The input vectors contain TB's and I_2 indexes for H and V polarizations and the three angle bins from 30º to 45º and the soil temperature from 0 to 7 cm from ECMWF IFS forecast. Therefore, input vectors have a total of 13 elements. All the 13 elements must be well-defined to train the NN and there must be a well-defined associated SM value. In addition, vectors with one of the I_2 indexes higher than 1 have been filtered-out. In addition, one fifth of the vectors in the training data base have been selected to have $\sim 3\,10^5$ vectors. The training method explained in [14] has been followed. A subset of 60% of those $\sim 3\,10^5$ vectors is used for the actual training, 20% is used for evaluation of the NN performances during the training and to avoid over-training, the final 20% is used to test the performances of the trained NN *a posteriori*. Gradient back-propagation and minimization with the Levemberg-Marquard algorithm has been used. One single hidden layer with 5 neurons has been used, as it has been shown in [14] that it is enough to capture the relationship in between the input data and the reference SM and one wants to keep the NN as simple as possible. No signs of overtraining have been found and the training has been stopped after 50 iterations when the mean squared difference is asymptotically approaching to a minimum.

Figure 4 shows a scatter plot obtained with the test subset of the training database. The SM retrieved with the NN is plotted versus the SMOS L2 SM. Note that the density scatter plot has a logarithmic scale, therefore, most of the points are within the inner densest region of the scatter plot. The linear regression is $SM_{NN} = 0.74\,SM_{L2} + 0.04$. The Pearson correlation R is 0.86, the standard deviation of the difference (STDD) is 0.068 and the Root Mean Square Error or Difference (RMSE) is also 0.068, which implies that there is not a significant bias in between both SM datasets.

The global correlation of the NN output and the reference data found in this study is lower than those obtained using the NRT prototype and L3 CATDS data. Several reasons can explain the differences. The L3 data are provided for ascending and descending semi-orbits and two independent NNs are trained for each data subset. In addition, the daily L3 data have been averaged when more than one measurement is available for a given day (basically at high latitudes) while in the operational NRT training database the higher latitudes can be over-represented. However, this effect has been evaluated using different latitudes ranges for the training dataset, or even equalizing the training database to ensure that an equal number of samples comes from all latitudes.

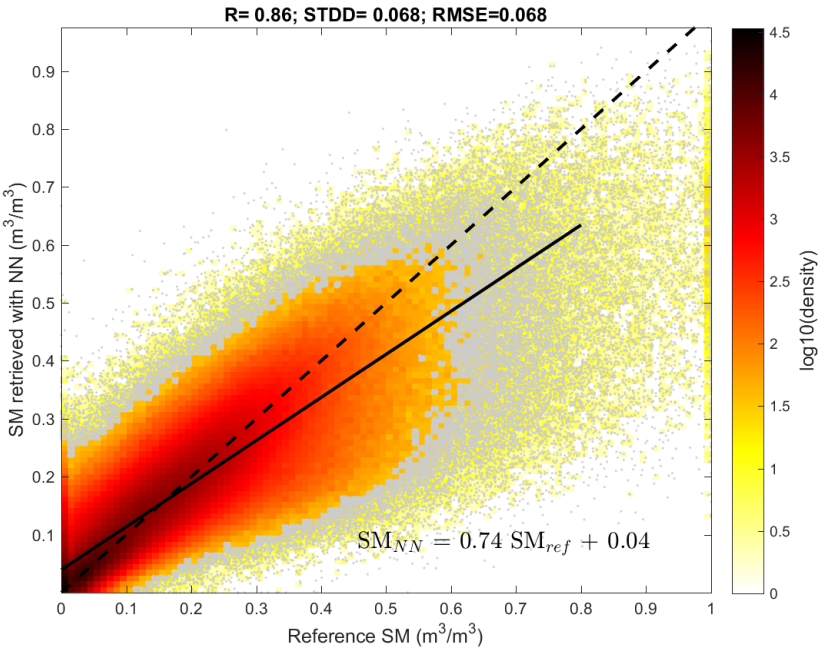

Figure 4: Scatter plot obtained with the test subset of the training database. The SM retrieved with the NN is plotted versus the SMOS L2 SM. Note that the density scatter plot has a logarithmic scale, therefore, most of the points are within the inner densest region of the scatter plot. The 1:1 line is shown with a dashed line and the regression line with a solid line. The linear regression is given in the bottom right of the plot. The Pearson correlation R, the standard deviation of the difference (STDD) and the Root Mean Square Error or Difference (RMSE) are given in the top of the figure.

However, these tests did not give a better training strategy.

Nevertheless, the results discussed above show that the operational NRT NN ability to capture the dynamics of the current L2 SM dataset is very good and the evaluation results discussed in Sect. 4 below confirm that the quality of the SM-NRT-NN product is as good as expected from the L3 prototype used to define the specifications of the operational product.

3.1 Neural network parameters

The NN architecture has been described in [14]. Here, the main characteristics of the NN and the error estimation have been summarized. The NN parameters are given in Tables 3 to 6.

Using the minimum and maximum values determined during the training phase (v_i^{min} and v_i^{max}, Table 3), the inputs have been re-normalized to obtain v_i^{norm}.

$$v_i^{norm} = -1 + 2\frac{v_i - v_i^{min}}{v_i^{max} - v_i^{min}}, \ \forall i = 1...13 \tag{1}$$

The normalized input, together with the first layer weights (W_{L1}, Table 4) and bias (B_{L1}, Table 5) are used to compute the first layer outputs v^{L1} as follows:

Table 3: Input vector elements and minimum and maximum values to be used to normalize the input vectors. The numbers in parenthesis in the second column is the center of the incidence angle bin in degrees.

	Elem.	v^{min}	v^{max}
$i=1$	$I_2^H(32.5)$	0.0	1.0
2	$I_2^H(37.5)$	0.0	1.0
3	$I_2^H(42.5)$	0.0	1.0
4	$I_2^V(32.5)$	0.0	1.0
5	$I_2^V(37.5)$	0.0	1.0
6	$I_2^V(42.5)$	0.0	1.0
7	$T_b^H(32.5)$	117.32	315.28
8	$T_b^H(37.5)$	111.79	319.13
9	$T_b^H(42.5)$	106.67	311.29
10	$T_b^V(32.5)$	137.47	344.98
11	$T_b^V(37.5)$	146.95	348.11
12	$T_b^V(42.5)$	151.58	345.30
$n_{in}=13$	$T_{(0-7cm)}$	274.00	334.13

Table 4: Weights of the first layer of neurons (W^{L1}) for each vector element i and neuron j.

	$i=1$	2	3	4	5
$j=1$	-0.251963	-0.221455	-0.200222	-0.183256	-0.122665
2	0.150564	0.013126	-0.191628	0.174118	0.001141
3	-0.063994	0.152765	0.186455	-0.720909	-0.518733
4	-0.421306	-0.499146	-0.405303	0.056720	-0.271721
$n_{L1}=5$	0.069875	0.052644	0.033591	0.076653	0.035225
	$i=7$	$i=8$	9	10	6
$j=1$	-0.178628	-0.581937	0.338913	1.967619	-1.987443
2	-0.193942	0.615439	1.433638	4.086621	-2.113233
3	-0.540105	-0.210301	-0.788601	3.002861	-3.769961
4	0.212869	-1.282507	-5.569205	-5.936806	5.805017
$n_{L1}=5$	-0.043341	-0.779192	0.163705	1.651020	-1.591232
	$i=11$	12	$n_{in}=13$		
$j=1$	-0.983539	-0.652086	0.874631		
2	-0.325188	-2.103513	-0.874681		
3	-4.703158	-7.372871	1.970066		
4	3.213482	5.677264	-1.035078		
$n_{L1}=5$	-0.545009	-0.149292	0.588369		

Table 5: Bias for the first layer of neurons (B^{L1}), weights for the second layer (W^{L2}), and bias for the second layer of neurons (B^{L2}).

	B^{L1}	W^{L2}	B^{L2}
$j=1$	-1.228212	-0.787173	-1.149465
2	-3.420649	-0.524800	...
3	3.376712	0.065983	...
4	1.652707	-0.206207	...
5	-0.077552	0.949027	...

Table 6: Parameters ($v^{L2}_{oldMax,newMin}$, ...) to apply to re-normalize the output of the NN (Eq. 4).

	old	new
Min	-1	0
Max	1	1

$$v^{L1}_j = \tanh(\sum_{i=1}^{13} W^{ij}_{L1}\, v^{norm}_i + B^j_{L1}),\ \forall j = 1...5 \tag{2}$$

The output of the second layer is computed from the first layer outputs, and the second layer weights W_{L2} and bias B_{L2} (both given in Table 5) as follows:

$$v^{L2} = \sum_{j=1}^{5} W^j_{L2} v^{L1}_j + B_{L2} \tag{3}$$

Finally, to obtain the NN output (v^{out}), the output of the second layer has to be re-normalized using the values v^{L2}_{min} and v^{L2}_{max} (Table 6) as follows:

$$v^{out} = v^{L2}_{newMin} + \frac{v^{L2}_{newMax} - v^{L2}_{newMin}}{v^{L2}_{oldMax} - v^{L2}_{oldMin}} (v^{L2} - v^{L2}_{oldMin}); \tag{4}$$

The NN output uncertainty has been estimated from the uncertainty in the inputs as explained in Sect 4.2 of [14].

4 SM-NRT-NN evaluation

4.1 Swath-level comparison to SMOS L2 SM

Figures 5 and 6 show the NRT-NN SM product and its associated uncertainty for two orbits of day 2012/05/27. In addition, the L2 SM and its associated uncertainty as given by the `DQX` parameter are also shown.

As discussed in [14], the swath width of the NRT-SM retrieval is somewhat narrower than the L2 SM one but both maps show similar spatial structures and numerical values. The uncertainties have similar numerical values as well, but the spatial patters are not necessarily similar for both products since the retrieval algorithms are different. Finally, it should be noted that the spacial coverage can be different for both products:

- the NRT-SM product can show circle-arc gaps when not all of the angle bins have a well defined TB value (left panels of Figs. 5 and 6). Future versions of the NRT-SM processor could include an interpolation strategy to estimate TBs values when there are gaps in the TB versus incidence angle profile.
- The NRT-SM global retrieval algorithm can provide a SM estimate even when the local minimization of the L2 algorithm does not converge. This can happen mainly in dense forest areas (right panels of Figs. 5 and 6).

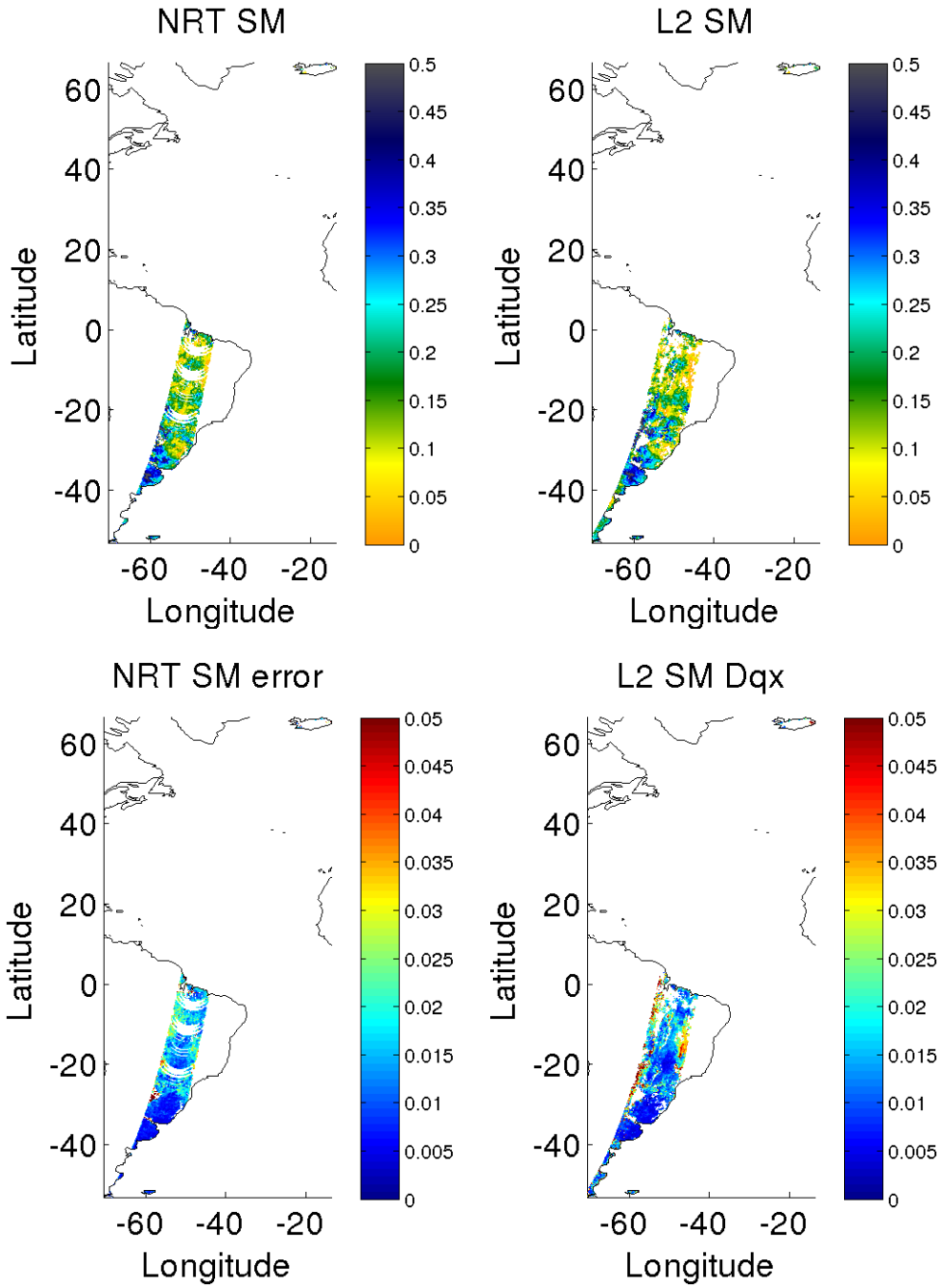

Figure 5: Comparison of the NRT-NN SM product (upper left) and the Level 2 SM (upper right) for 20h42m21s orbit of day 2012/05/27. The corresponding NRT-NN uncertainty is shown in the lower left panel, while the L2 SM uncertainy is shown in the lower right panel as given by the `Dqx` *parameter.*

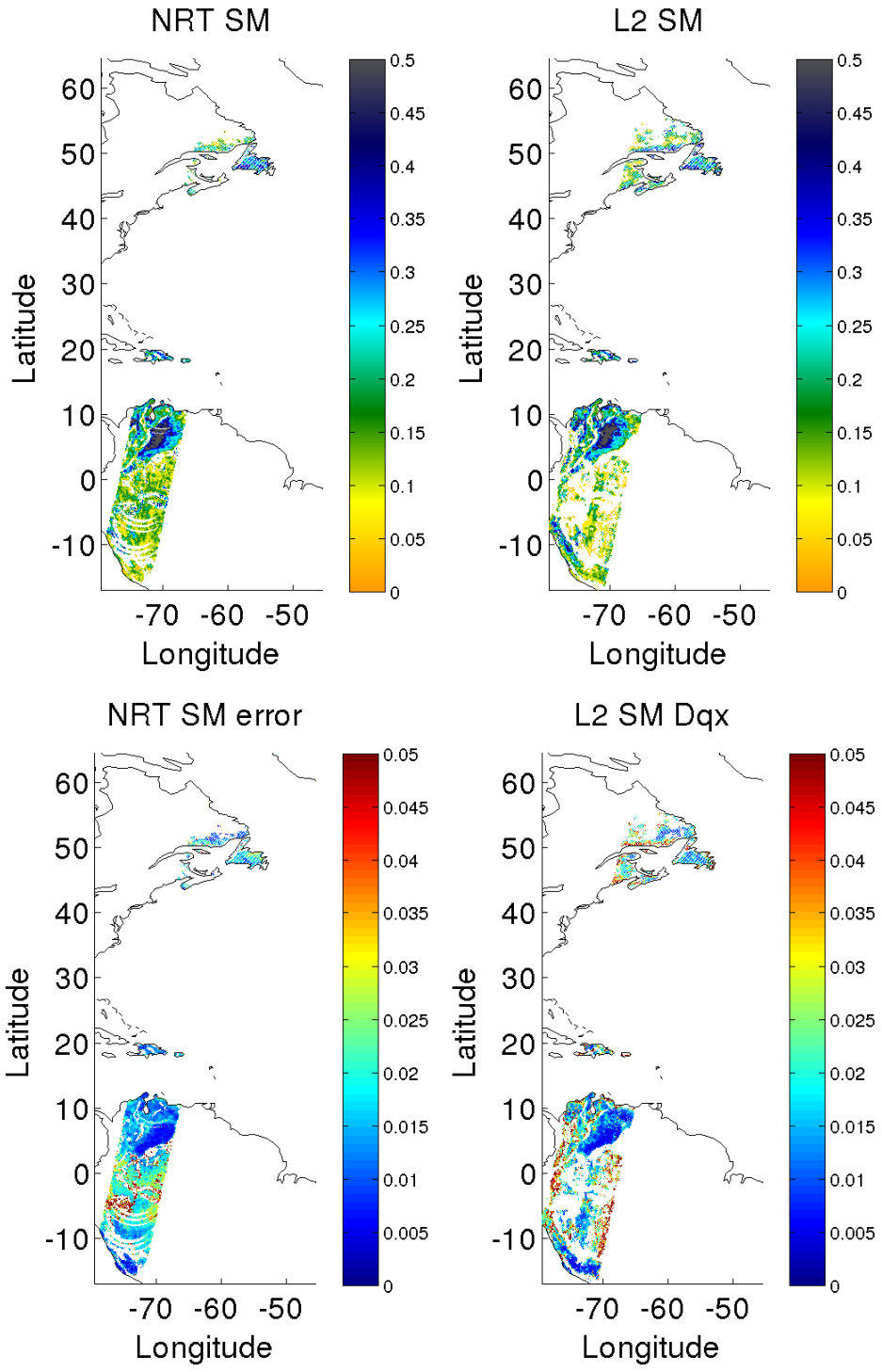

Figure 6: Same as Fig. 5 but for 22h22m32s orbit of day 2012/05/27.

4.2 Global and long-term comparison to SMOS L2 SM

The SM-NRT-NN product has also been compared to the SMOS L2 SM product globally and over a long period. The NRT SM processor has been applied to NRT TBs and compared to SMOS L2 SM obtained with version version 6.20 of the SMOS operational processor from 15/May/2015 to 25/November/2015. Figures 7 to 9 show maps and histograms computed from local (all DGGs) statistical metrics obtained over that period. The typical number of points with both NRT-SM-NN and SM-L2 in that period is ~ 110 (upper panel of Fig. 7). The correlation of both products is very high (> 0.7) over a large part of North-America, the southernmost part of South-America, the Iberian peninsula, the Sahel and South-Africa, Australia and parts of central Eurasia. The correlation in significantly lower over forest (both tropical and boreal) and in deserts such as the Sahara, where the variance is low and driven by the noise. In conclusion, both products show similar dynamics over large parts of the Globe. The Bias map (upper panel of Fig. 8) show that the SM-NRT-NN product shows a tendency to underestimate the L2-SM dataset, which is a expected behaviour as it has been obtained using a regression technique and extreme values are under-represented in the reference dataset. The most significant effect of the bias is to increase the RMSD with respect to the STDD in parts of Europe and Canada. However, one should note that both the RMSD and the STDD are lower than 0.04 m^3/m^3 over most of the Globe (all except the reddish regions in the two lower panels of Fig. 8, see also the positions of the peak of the histograms.)

Finally, Fig. 9 shows the mean of the SM-NRT-NN and SM-L2 over the period of the study. Both maps show an overall excellent agreement, although it is possible to appreciate the negative bias in the NRT-SM-NN product in some regions. In any case, note that it has been preferred to compare the mean of both products instead of the median to be more sensitive to the possible underestimation of extreme values in the SM-L2 dataset.

The same kind of evaluation but using reprocessed data over a period of two years is briefly discussed in Appendix A. The results are comparable to those obtained with operational data in a shorter period discussed here but the Pearson correlation map shows higher values over larger regions of the Globe.

4.3 Evaluation with respect to *in situ* measurements

The SM-NRT-NN product has been evaluated against *in situ* measurements from the SCAN [16] and USCRN [4] networks. This networks of *in situ* measurements have been extensively used for the validation of remote sensing data [3, 13, 1, 11, 2, 8]. The *in situ* data have been obtained directly from the teams operating both networks. After a quality inspection of the data, the sites listed in Appendix A have been selected for the current study. Some individual time series are also briefly discussed in Appendix A.

Several quality metrics have been computed site per site independently for the SM-NRT-NN and the SM-L2 products. Table 7 summarizes the results in the form of averages over all the sites (for the Pearson correlation also the median value is given).

The mean number of points in the time series from May 2015 to November 2015 is 186 for the SM-L2 product while is only half of that value for the SM-NRT-NN product. The reason is a larger revisit time for the SM-NRT-NN product due to the somewhat narrower swaths of the retrieval and the lack of retrievals if not the 6 TBs are well defined for both polarizations and the three angle bins from 30^o to 45^o. It is noteworthy that the current implementation of the SM-NRT-NN processor do not perform any interpolation of the TB versus incidence angle profiles. Both SMOS product shows the very close mean bias with respect to the *in situ* measurements, while the mean STDD and RMSD are slightly lower for the SM-NRT-NN product.

In order to get further insight into the intrinsic quality differences of both datasets, the same statistics have been computed but only keeping times for which both SMOS products are retrieved. The results are shown in Table 8. The differences in the evaluation of both products decreases, but the NRT product still shows a larger

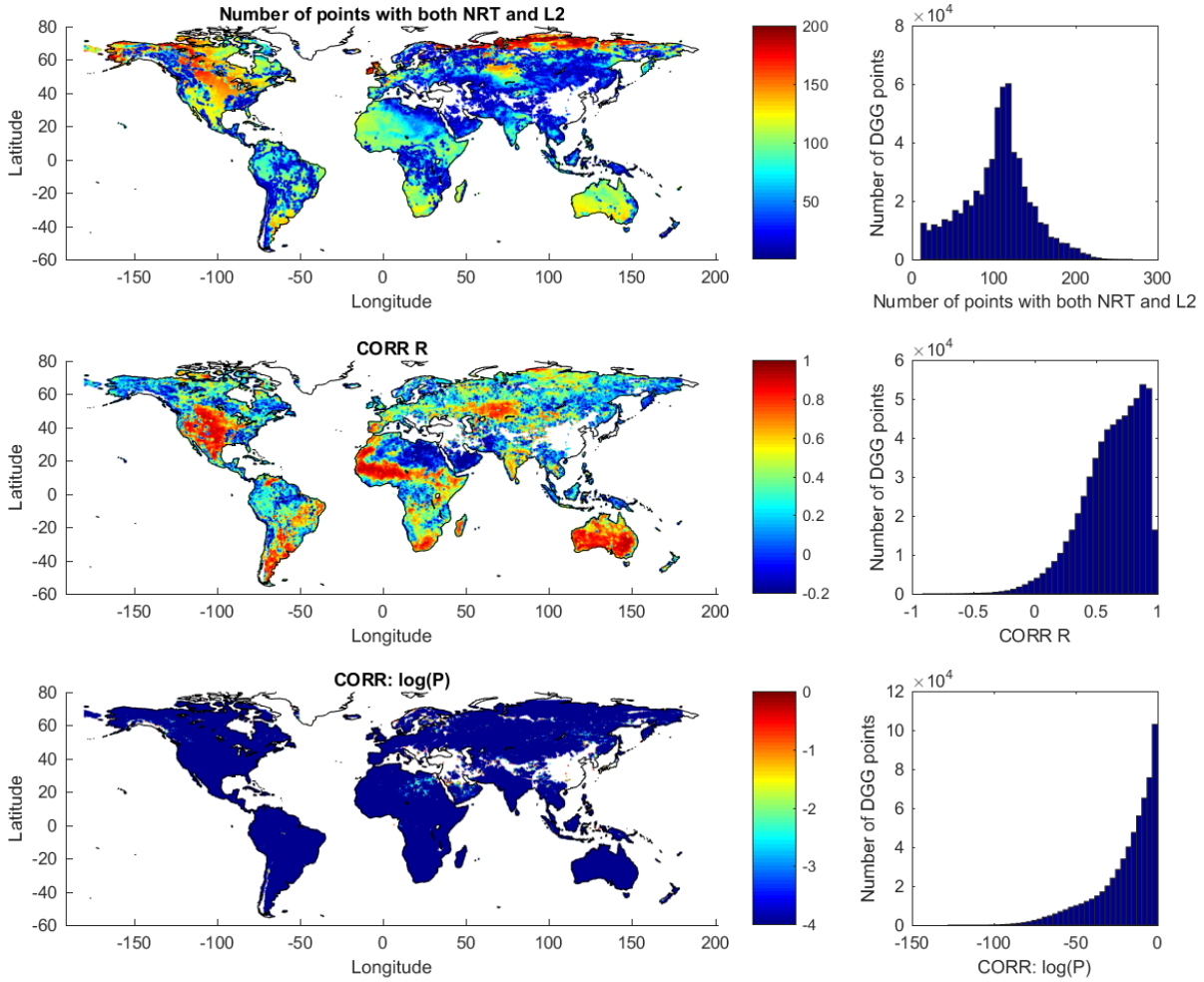

Figure 7: From top to bottom: maps (left) and histograms (right) of the number of points (with both SM NRT and SM L2), the Pearson correlation of NRT SM with respect to L2 SM and the probability of have a given correlation by chance (in logarithmic scale) for each DGG point.

Table 7: Comparison to in situ measurements over the USCRN and SCAN networks. The columns are: the SM product, the mean number of points in the time series, the mean and median Pearson correlation with respect to in situ measurements, the mean bias (mean in situ SM minus mean SMOS SM), the RMS and STD of the difference time series averaged over all sites, and the Anomaly Correlation Coefficient [3]. The statistics have been computed independently for the SM-NRT-NN and the SM-L2 product. The number of SM retrievals is, on average, larger for the SM-L2. The number of sites with available data for the evaluation is 144 for SM-L2 and 155 for SM-NRT-NN.

SM	Mean N_{pts}	Mean R	Median R	mean Bias	mean RMSD	Mean STDD	mean ACC
L2	186	0.63	0.64	0.035	0.100	0.065	0.56
NRT	94	0.70	0.71	0.036	0.095	0.058	0.48

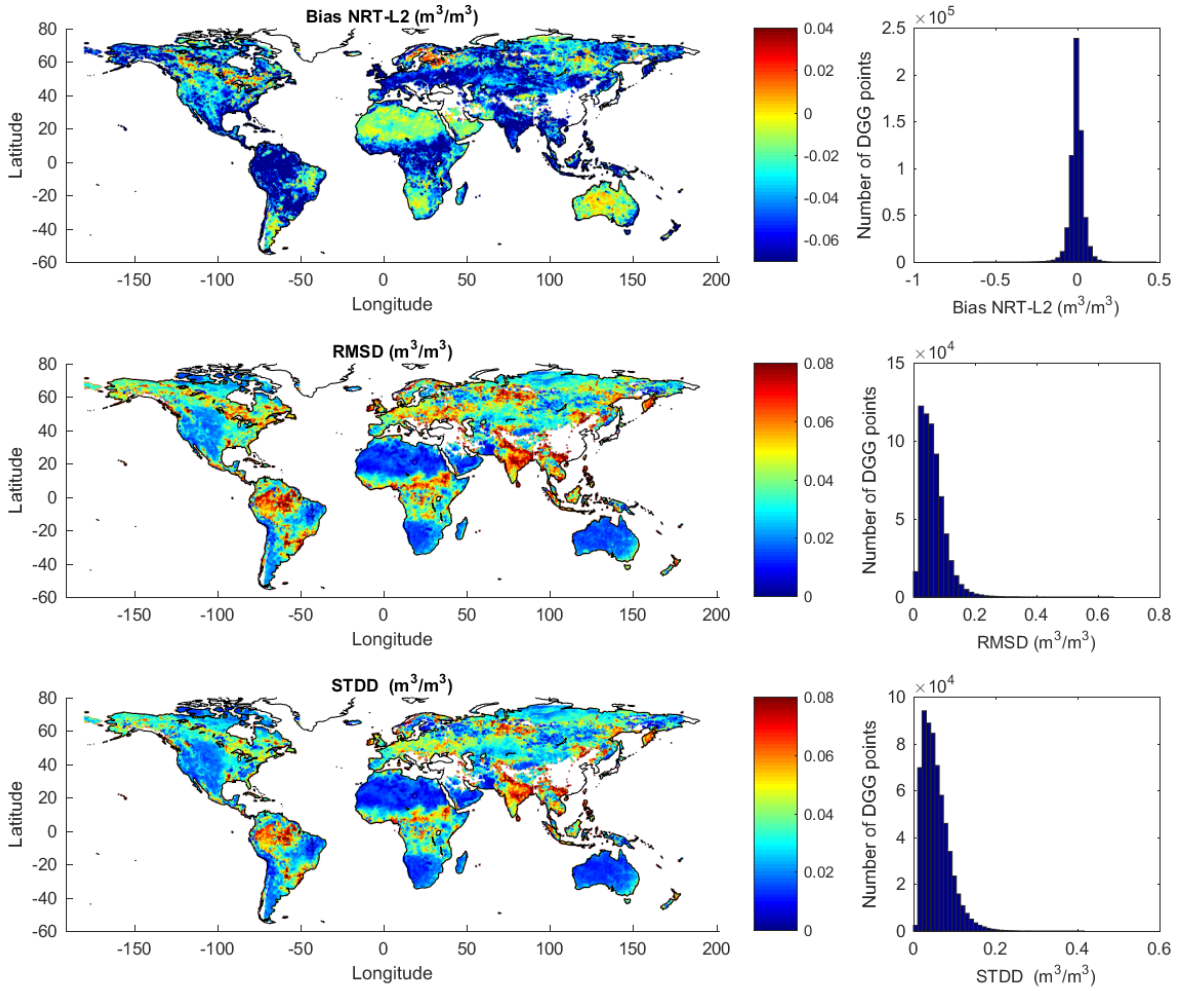

Figure 8: From top to bottom: maps (left) and histograms (right) of the bias (mean SM NRT minus mean SM L2), the RMS and the STD of the difference of NRT SM with respect to L2 SM for each DGG point.

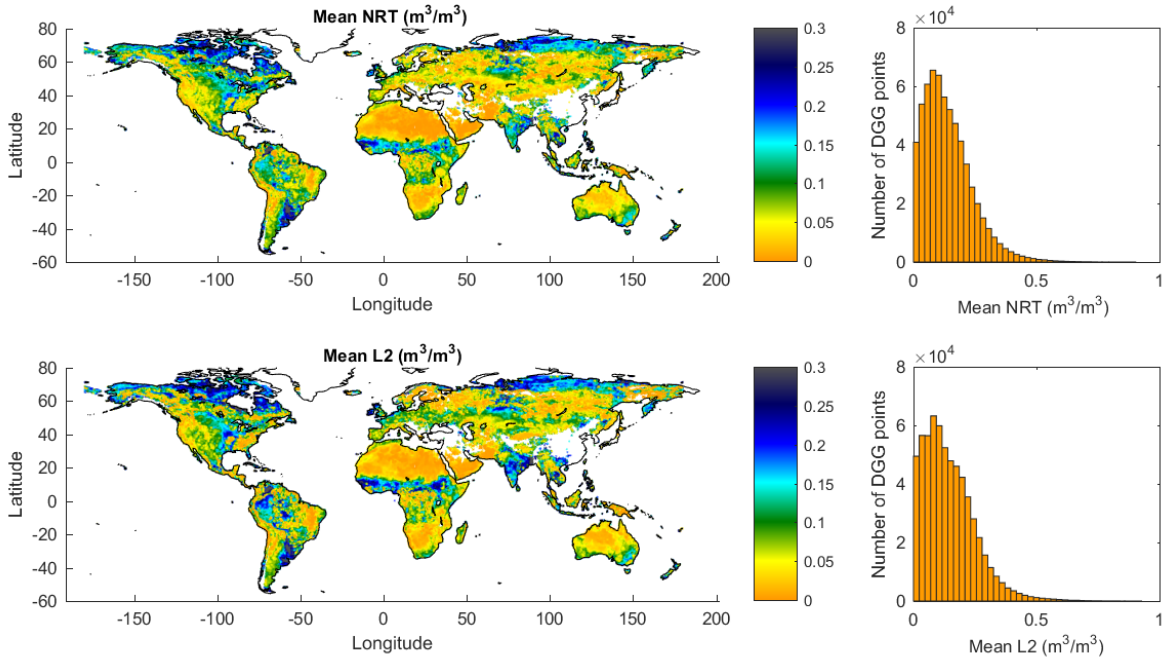

Figure 9: From top to bottom: maps (left) and histograms (right) of the mean NRT SM and mean L2 SM for each DGG point.

Table 8: Same as Table 7 but the statistics have been computed only with times for which both the SM-NRT-NN and the SM-L2 products are available. The total number of sites selected is 127. Table 9 shows the coordinates of all the sites.

SM	Mean N_{pts}	Mean R	Median R	mean Bias	mean RMSD	Mean STDD	mean ACC
L2	88.	0.67	0.69	0.026	0.092	0.062	0.59
NRT	88.	0.71	0.72	0.031	0.091	0.056	0.56

correlation and lower STDD with respect to *in situ* measurements than the L2 product. This result is in perfect agreement with the results obtained with the NRT prototype implemented in the first phase of this project [14].

Since the mean or median values alone do not show the full picture of the evaluation for more than 100 sites, Figure 10 shows boxplots for the Pearson correlation coefficient R ("CC" in the figure), bias, RMSD, STDD ("ubRMSD" in the figure), ACC [Anomaly Correlation Coefficient, see 3] and the number of points in the time series (equal for SM-L2 and SM-NRT-NN in this case).

As expected, there is a large variation from one site to another. The Bias and STDD distribution are similar for both products, while higher values of RMSD are found for some NRT time series. In contrast, the correlation is as high as almost 1 for some sites both for the SM-NRT-NN and SM-L2 (the maximum is slightly higher for the later). Interestingly, the lower values of the distribution of the correlation are higher for the NRT product.

Finally, Fig. 11 shows scatter plots of the correlation and the anomaly correlation coefficient for both products taking into account the respective confidence intervals. For most of the sites, both products show the same statistics with respect to the *in situ* measurements and globally, the scatter plot points lie close to the 1:1 line.

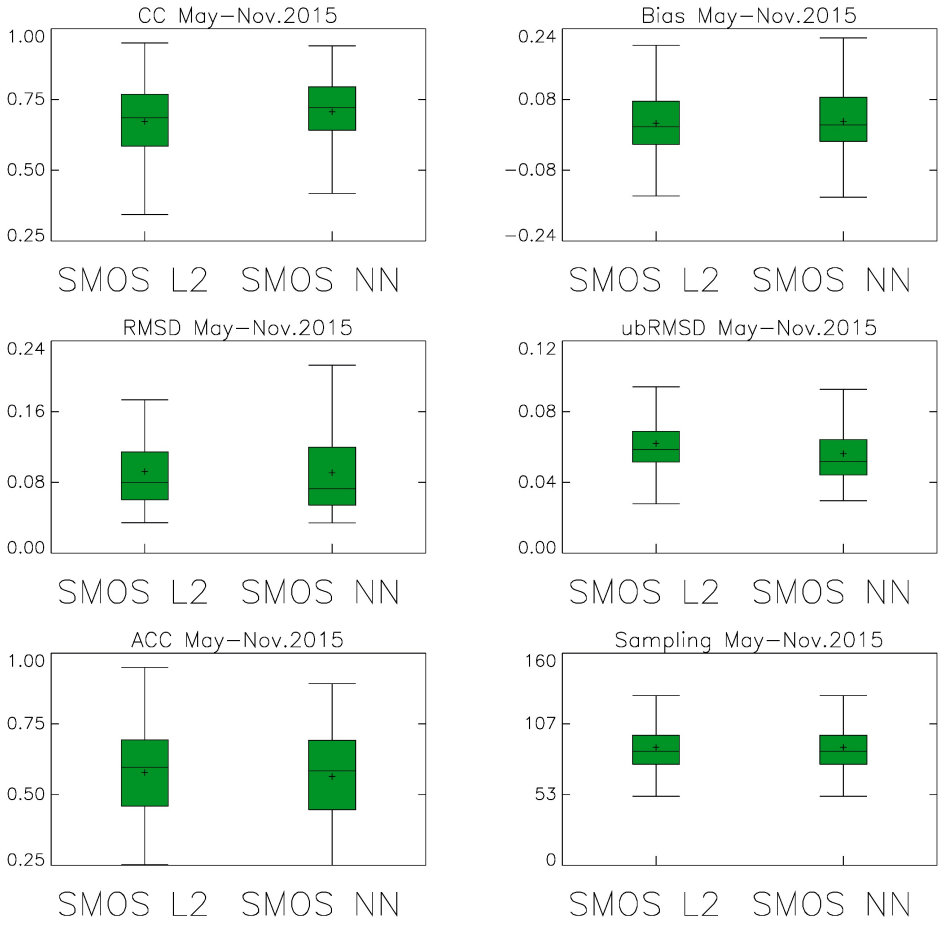

Figure 10: Boxplots for the Pearson correlation coefficient (CC), Bias (mean in situ minus mean SMOS SM), RMSD, STDD (ubRMSD), anomaly correlation coefficient (ACC) and number of points per time series (sampling) the SM-L2 (SMOS L2) and SM-NRT-NN (SMOS NN) products in comparison to in situ measurements. The box contains the middle 50% of the data, the central bar represents the median value of the distribution. The upper edge (hinge) of the box indicates the 75th percentile of the data set (q_3), and the lower hinge indicates the 25th percentile (q_1). The mean values are also shown as black crosses. The upper and lower bars represent the minimum and maximum values of the distribution excluding outliers. Points are considered as outliers if they are larger than $q_3 + 1.5(q_3 - q_1)$ or smaller than $q_1 - 1.5(q_3 - q_1)$.

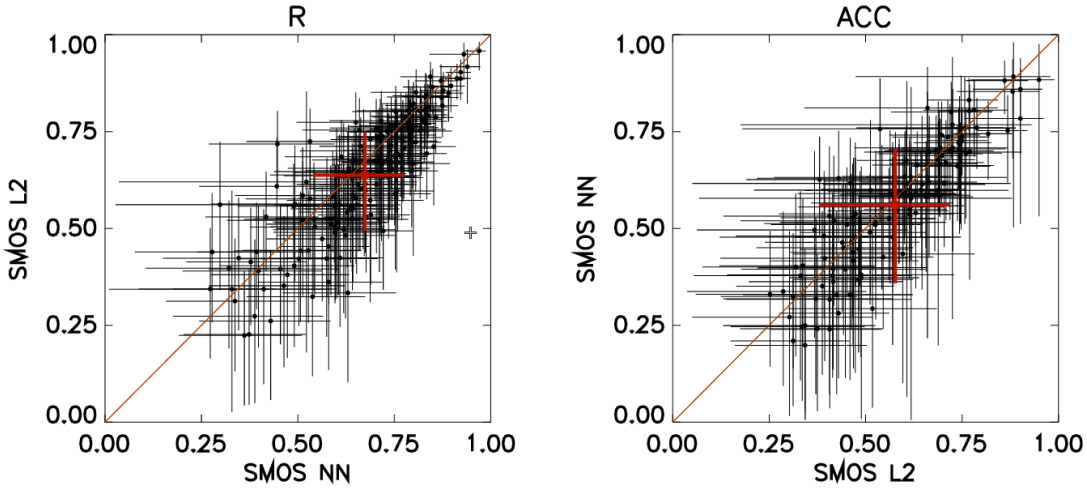

Figure 11: Scatter plots showing the Pearson correlation coefficient (R) and correlations on anomaly time-series (ACC) for the SM-NRT-NN (SMOS NN) dataset against in situ measurements versus R and ACC for the L2-SM (SMOS L2) dataset against in situ measurements. The errorbars account for the 95% confidence intervals. The red symbols represents an averaged value.

5 Summary and perspectives

This document discusses the implementation of the SMOS NRT SM processor using the neural network algorithm described in [14]. The NRT TBs in BUFR format are filtered and processed to give to the neural network the input information in the required format. In addition, ECMWF forecast fields of soil temperature and snow cover are interpolated to complement the input data and filter out some retrievals.

The SM-NRT-NN product has been evaluated with respect to the original SM-L2 product in orbit basis and globally using several months of data. The NRT SM product compares well with the L2 product although it shows a negative bias for the highest SM values.

The SM-NRT-NN product has also been evaluated with respect to *in situ* measurements of SM over the SCAN and USCRN networks. The NRT product show similar performances than the L2 product. The mean and median correlation and STDD are higher and lower, respectively, than those found for the L2 product, in agreement with previous results with the NRT prototype compared to SMOS CATDS L3 SM data.

In summary, the SM-NRT-NN product is similar to the current operational one but available in near-real-time. NRT TBs are received by ECMWF in less than three hours after sensing. The NRT SM production takes on average 15 minutes (the arrival of new NRT TBs is checked every 30 minutes and the actual NRT SM production takes a few minutes).

5.1 Maintenance and potential perspectives of development

This document describes the first version of the SM-NRT-NN processor. It is important to note that code maintenance will be needed to:

- Keep on with possible updates in input NRT brightness temperatures and with possible updates in libraries and third-party packages dependencies. Some code optimization and reorganization would ease maintenance and portability.

- Update brightness temperatures and SM-L2 maximum and minimum tables. Retrain the NN using an extended training dataset including last year SM-L2 data. This can be done typically once per year.
- Re-train the NN if a new version of SMOS SM-L2 is released. In this case, reprocessing the whole SM-NRT-NN dataset with the new NN would be desirable.

In addition, future versions of the processor could include the following algorithmic developments:

- NRT angle-binned HV brightness temperatures: Relaxing conditions of full interpolation, by investigating the impact of allowing extrapolation and interpolation between gaps of more than 2 integration times.
- Identify ascending/descending semi-orbits in the NRT brightness temperatures data and investigate dedicated NNs for each type of semi-orbits.
- Reprocessing the whole SM-NRT-NN dataset would be desirable in case of new algorithmic developpements.

Finally, of course, it would be good to include feedback from users in the following versions of the SM-NRT-NN processor.

References

[1] A. Al Bitar, D. Leroux, Y. H Kerr, O. Merlin, P. Richaume, A. Sahoo, and E.F. Wood. Evaluation of SMOS Soil Moisture Products Over Continental US Using the SCAN/SNOTEL Network. *Geoscience and Remote Sensing, IEEE Transactions on*, 50:1572–1586, May 2012.

[2] C. Albergel, P. de Rosnay, C. Gruhier, J. Muñoz-Sabater, S. Hasenauer, L. Isaksen, Y. Kerr, and W. Wagner. Evaluation of remotely sensed and modelled soil moisture products using global ground-based in situ observations. *Remote Sensing of Environment*, 118:215–226, 2012.

[3] C. Albergel, C. Rüdiger, D. Carrer, J.C. Calvet, N. Fritz, V. Naeimi, Z. Bartalis, and S. Hasenauer. An evaluation of ASCAT surface soil moisture products with in-situ observations in Southwestern France. *Hydrology and Earth System Sciences*, 13(2):115–124, February 2009.

[4] J.E. Bell, M.A. Palecki, C.B. Baker, W.G. Collins, J.H. Lawrimore, R.D. Leeper, M.E. Hall, J. Kochendorfer, T.P. Meyers, T. Wilson, and H.J. Diamond. U.s. climate reference network soil moisture and temperature observations. *J. Hydrometeorol.*, 14:977–988, 2013.

[5] L. Berthon, A. Mialon, F. Cabot, A. Al Bitar, P. Richaume, Y. Kerr, D. Leroux, S. Bircher, H. Lawrence, A. Quesney, and E. Jacquette. CATDS Level 3 data product description: Soil Moisture and Brightness Temperature part. *CESBIO report SO-TN-CB-CA-0001*, pages 1–38, February 2013.

[6] P. de Rosnay, M. Dragosavac, M. Drusch, A. Gutiérrez, M. Rodríguez López, N. Wright, J. Muñoz Sabater, and Crapolicchio R. SMOS NRT BUFR specification. SMOS-NRT-BUFR-ECMWF - v2.0, ECMWF, 2012.

[7] A. Gutierrez and A. Canales Molina. SMOS NRT Product Format Specification. Technical Report SO-ID-DMS-GS-0002, DEIMOS Space, 2010.

[8] Y. H. Kerr, A. Al-Yaari, N. Rodriguez-Fernandez, M. Parrens, B. Molero, D. Leroux, S. Bircher, A. Mahmoodi, A. Mialon1, P. Richaume, S. Delwart, A. Al Bitar, T. Pellarin, R. Bindlish, T. J. Jackson, C. Rudiger, P. Waldteufel, S. Mecklenburg, and J.-P. Wigneron. Overview of SMOS performances in term of global soil moisture monitoring after 5 years in operation. *Remote Sensing of Environment*, submitted, 2016.

[9] Y.H. Kerr, P. Waldteufel, P. Richaume, I. Davenport, P. Ferrazzoli, and J.P. Wigneron. SMOS level 2 processor soil moisture algorithm theoretical basis document (ATBD). *SM-ESL (CBSA), CESBIO, Toulouse, SO-TN-ESL-SM-GS-0001-3.g, SO-TN-ARR-L2PP-0037-3.6*, 2011.

[10] Y.H. Kerr, P. Waldteufel, P. Richaume, J.P. Wigneron, P. Ferrazzoli, A. Mahmoodi, A. Al Bitar, F. Cabot, C. Gruhier, S. Juglea, D. Leroux, A. Mialon, and S. Delwart. The SMOS Soil Moisture Retrieval Algorithm. *IEEE Transactions on Geoscience and Remote Sensing*, 50:1384–1403, May 2012.

[11] J. Kolassa, F. Aires, J. Polcher, C. Prigent, C. Jimenez, and J. M. Pereira. Soil moisture retrieval from multi-instrument observations: Information content analysis and retrieval methodology. *Journal of Geophysical Research: Atmospheres*, 118(10):4847–4859, 2013.

[12] A. Quesney. Data Processing Model: traitements L3TB. Technical Report CAT-DPM-CTL3TB-00061-CG, CapGemini Sud / ACRI-ST, 2011.

[13] N. J. Rodríguez-Fernández, F. Aires, P. Richaume, Y. H. Kerr, C. Prigent, J. Kolassa, F. Cabot, C. Jiménez, A. Mahmoodi, and M. Drusch. Soil moisture retrieval using neural networks: application to SMOS. *IEEE Transactions on Geoscience and Remote Sensing*, Volume:PP , Issue: 99, 2015.

[14] N. J. Rodríguez-Fernández, P. Richaume, J. Muñoz-Sabater, P. de Rosnay, and Y. H. Kerr. SMOS Near-Real-Time Soil Moisture processor. Part1: Neural Networks evaluation and algorithm description. Technical Report SMOS Ground Segment SO-TN-CB-GS-049, CESBIO, Toulouse, France, 2015.

[15] K. Sahr, D. White, and Kimerling A. J. Geodesic discrete global grid systems cartography. *Cartography and Geographic Information Science, 30(2)*, 30(2):121–134, 2003.

[16] G.L. Schaefer, M.H. Cosh, and T.J. Jackson. The USDA natural resources conservation service soil climate analysis network (SCAN). *Journal of Atmospheric and Oceanic Technology*, 24(12):2073–2077, 2007.

Acknowledgements

This work is funded under the ESA-ESRIN contract number 4000101703/10/NL/FF/fk. We would to thank Matthias Drusch, Susanne Mecklenburg and Antonio de la Fuente (all ESA staff) for their useful comments at different stages of this project.

A Evaluation supplement material

A.1 Global analysis

Since, the period available with operational v6.20 data is relatively short, covering not even a whole year, the same kind of global analysis discussed in Sect. 4 has been done applying the SM-NRT-NN processor to NRT TBs reprocessed with the same parameters of current operational NRT TBs. The period of the study was June 2010 - June 2012. One should bear in mind that for this analysis the filtering was a bit different with respect to that described above in this document. For this long period study all the TBs with a RFI-related bit on (bits 1, 4 or 9), have been filtered out.

The NRT-SM-NN have been compared to v6.20 L2 SM. Figures 12 and 13 show The results are comparable to those obtained with operational data in a shorter period discussed above. However, it is noteworthy that the Pearson correlation map (upper panel of Fig. 12 shows higher values over larger regions of the Globe, which shows that the temporal dynamics of both products are actually in better agreement when a longer time period is considered.

A.2 Time series

Table 9 shows the coordinates of all the sites used for the evaluation against *in situ* measurements.

Figures 14 and 15 show examples of the three time series for some sites selected randomly. As expected, for some sites the two SMOS products are very different to the *in situ* measurements. This is most likely due to the different resultion of the remote sensing measurement with respect to an *in situ* point measurement. Different sensing depths can also explain the differences for some sites. It is also possible to see sites for which the L2 product seem to be closer to the *in situ* measurement and sites for which the NRT product is closer.

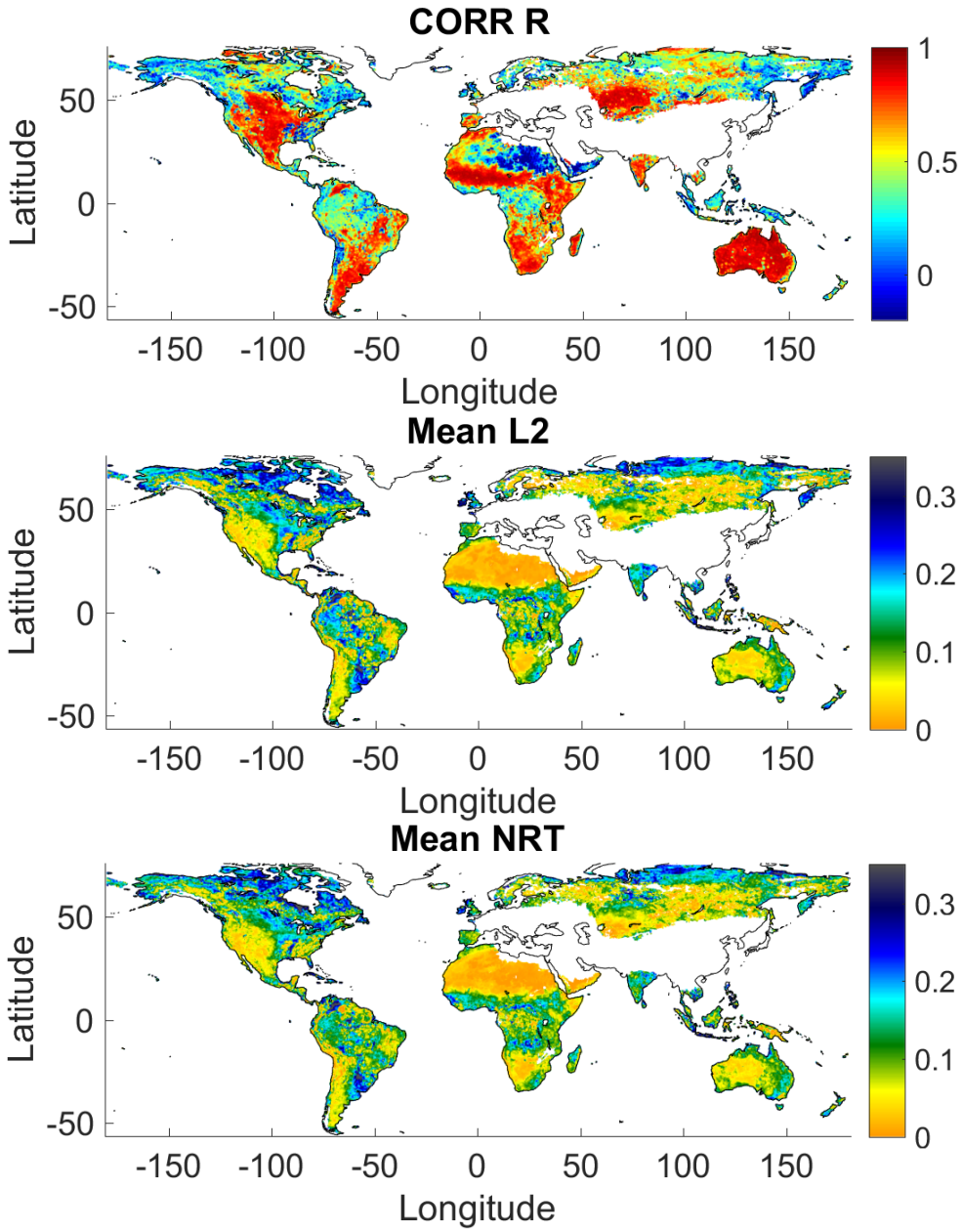

Figure 12: From top to bottom: (i) Pearson correlation (adimensional) in between SM-NRT-SM and SM-L2. (ii) Local mean value of SM-L2 (m^3/m^3). (iii) Local mean value of SM-NRT-NN (m^3/m^3). These maps have been computed comparing the NRT-SM processor output applied to reprocessed TBs and the SM-L2 reprocessed v6.20 data from June 2010 to June 2012.

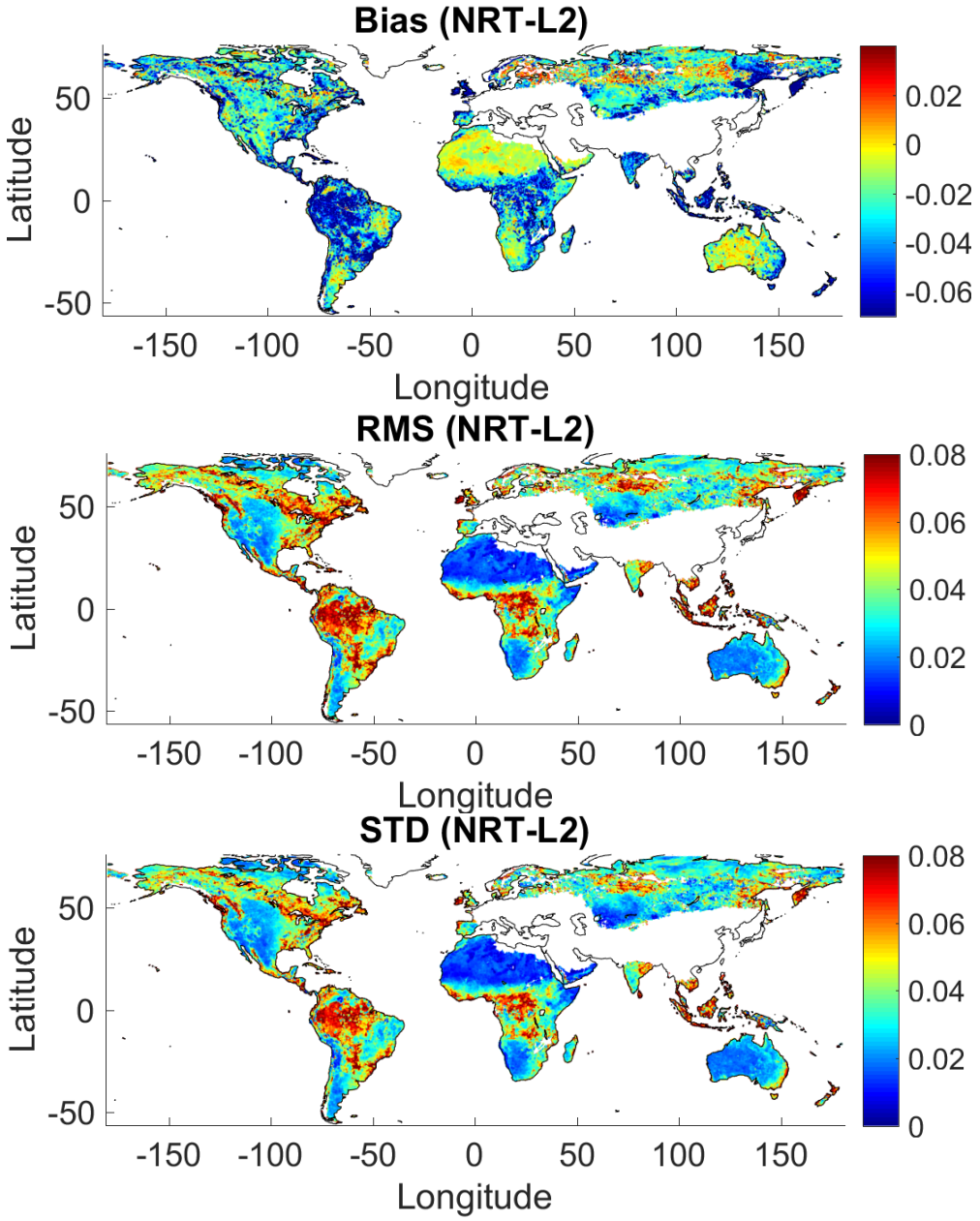

Figure 13: From top to bottom: (i) Bias of the SM-NRT-NN product with respect to the SM-L2 product: mean of SM-NRT-NN minus mean SM-L2 (m^3/m^3). (ii) Root mean square of the difference (m^3/m^3). (iii) Standard deviation of the difference (m^3/m^3). These maps have been computed comparing the NRT-SM processor output applied to reprocessed TBs and the SM-L2 reprocessed v6.20 data from June 2010 to June 2012.

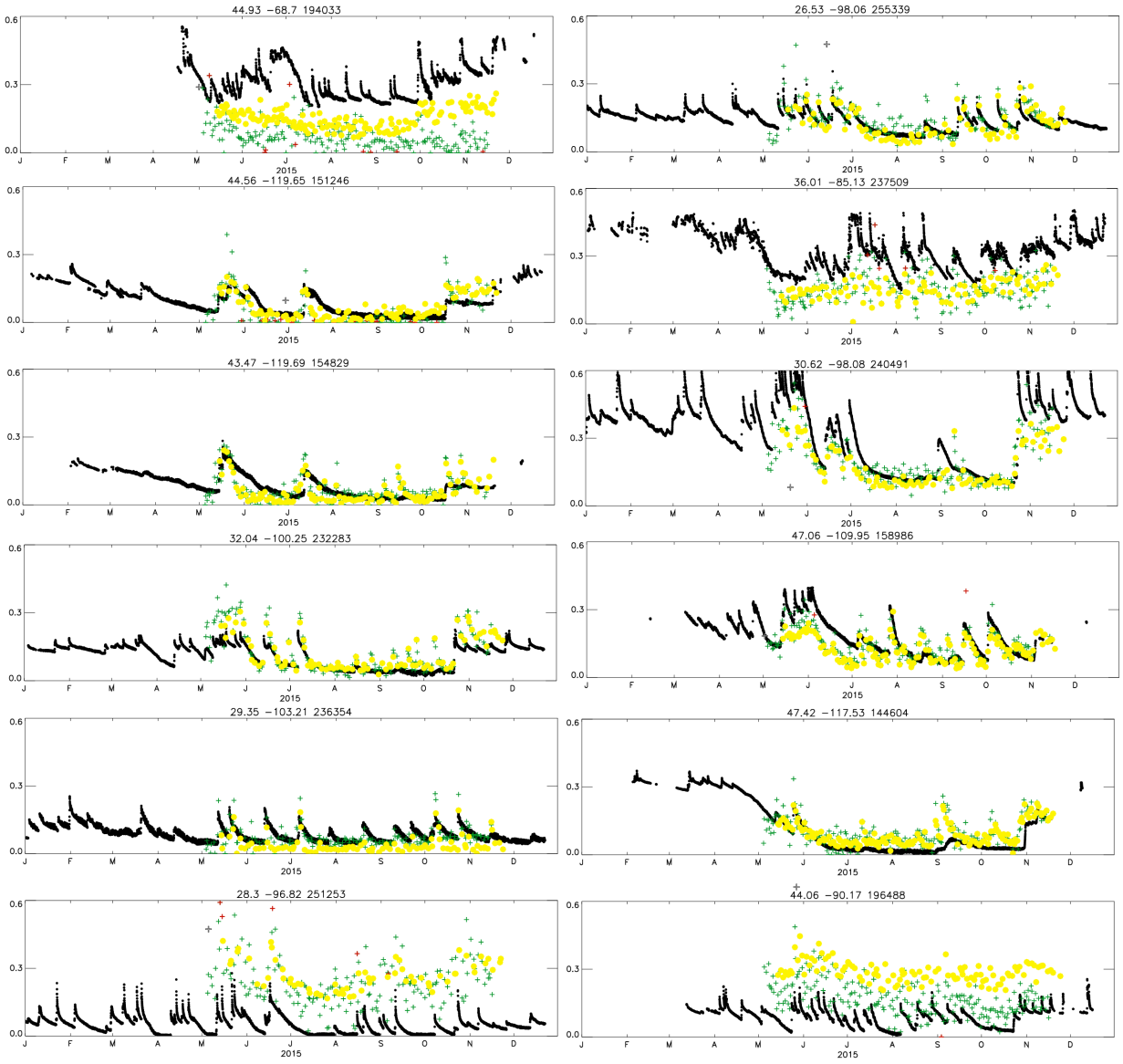

Figure 14: Soil moisture time series for a selection of in situ measurement sites. Yellow dots: SM-NRT-NN. Green crosses: L2. Black dot: in situ measurements.

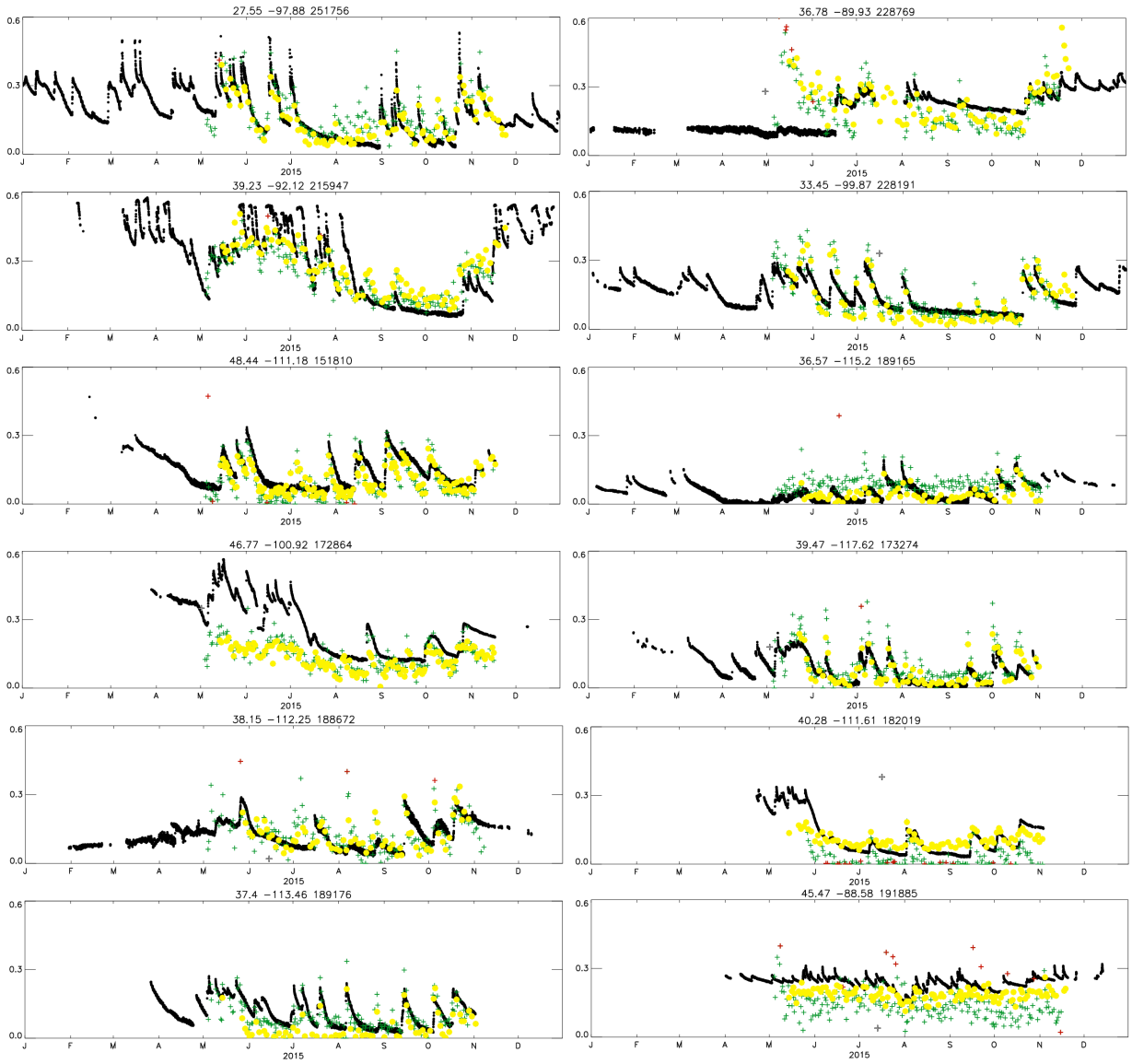

Figure 15: Soil moisture time series for a selection of in situ measurement sites. Yellow dots: SM-NRT-NN. Green crosses: L2. Black dot: in situ measurements.

Table 9: List of the sites used for the simultaneous evaluation of SM-L2 and SM-NRT-NN (127 sites).

lat	lon	lat	lon	lat	lon
deg	deg	deg	deg	deg	deg
31.59	-110.51	44.02	-100.35	34.36	-106.69
32.24	-111.17	43.73	-96.62	36.37	-115.83
35.76	-112.34	30.62	-98.08	40.03	-118.18
38.54	-107.69	32.04	-100.25	39.47	-117.62
27.15	-81.37	26.53	-98.06	42.88	-77.03
31.19	-84.45	31.62	-102.81	34.3	-79.73
31.31	-84.47	33.96	-102.77	43.74	-96.61
43.2	-116.75	31.78	-95.72	35.07	-86.89
40.05	-88.37	29.35	-103.21	27.55	-97.88
41.84	-88.85	41.62	-112.54	33.45	-99.87
39.1	-96.61	44.06	-90.17	30.08	-95.98
38.87	-100.96	44.52	-104.44	31.48	-96.88
38.09	-84.75	42.43	-95.77	31.55	-100.51
30.09	-91.87	40.32	-89.9	32.25	-98.2
46.96	-67.88	37.13	-97.09	29.36	-100.25
48.31	-95.87	39.7	-96.16	34.02	-99.25
39.87	-93.15	39.05	-95.19	26.16	-97.96
37.43	-94.58	39.79	-99.33	37.67	-109.36
45.16	-113.01	37.19	-86.04	41.94	-112.43
46.88	-110.29	45.42	-93.95	41.34	-111.19
48.74	-113.43	47.72	-96.27	41.82	-111.98
48.31	-105.1	39.23	-92.12	38.15	-112.25
48.49	-105.21	38.87	-94.03	37.63	-113.64
46.77	-99.48	37.99	-94.04	40.59	-112.39
46.89	-103.38	40.25	-93.72	41.78	-113.82
48.97	-102.17	33.38	-90.65	38.59	-113.75
42.42	-103.74	34.25	-89.87	38.31	-109.24
40.7	-96.85	34.23	-89.9	40.18	-110.3
40.85	-96.57	32.86	-90.52	37.45	-109.34
42.07	-101.44	33.97	-90.9	40.37	-110.41
32.61	-106.74	33.62	-91.1	37.87	-112.43
34.36	-106.89	33.09	-90.51	41.77	-113.26
39.01	-114.21	33.63	-88.77	39.53	-110.81
36.62	-116.02	34.68	-90.42	40.38	-109.36
41.79	-73.74	34.07	-90.35	37.51	-111.26
36.6	-101.59	48.48	-109.8	38.01	-109.13
36.57	-101.61	47.52	-107.13	36.63	-80.13
36.12	-97.09	47.06	-109.95	42.71	-90.39
36.13	-97.11	45.8	-111.59	40.36	-116.86
44.56	-119.65	48.44	-111.18	41.79	-111.55
33.36	-81.33	46.77	-100.92	40.28	-111.61
45.71	-99.13	34.25	-105.42	...	...
45.52	-103.3	34.77	-106.76	...	...