Read-me-first note for the release of the SMOS Level 2 Soil Moisture Near Real Time Neural Network (L2-SM-NRT-NN) data product version 300	
Processor version	Level 2 Soil Moisture Near Real Time Neural Network v300
Release date by ESA	2nd August 2021
Further information	Details on the neural network design, training and the processor can be found in: **[RD1]** Rodríguez-Fernández, N. J., et al., 2021: SMOS Near-Real-Time soil moisture processor version 300. Neural network design and first evaluation results, CESBIO-SMOS report SO-TN-CB-GS-0100 v1.0, 29.06.2021 Report available here **[RD2]** Rodríguez-Fernández, N. J., et al., 2016: SMOS Near-Real-Time soil moisture processor. Part 1: Neural network evaluation and algorithm description, CESBIO-SMOS report SO-TN-CB-GS-0049 v1.6, 23.03.2016 **[RD3]** Muñoz-Sabater, J., et al., 2016: SMOS Near-Real-Time soil moisture processor: Operational chain and evaluation, ECMWF TR2, WP 4020, 01/2016 **[RD4]** Rodríguez-Fernández, N. J., Muñoz-Sabater, J.; Richaume, P.; Albergel, C.; de Rosnay, P. & Kerr, Y. H. Evaluation of the SMOS Near-Real-Time soil moisture, Hydrology and Earth System Sciences, 2017, 21, 5201-5216 **[RD5]** Rodríguez-Fernández, N. J., et al., 2021: SMOS Near-Real-Time soil moisture processor trained on Level 2 v650 soil moisture data. Neural network weights and evaluation report, CESBIO-SMOS report SO-TN-CB-GS-0060 v1.0, 16.05.2018 **[RD6]** Rodríguez-Fernández, N. J., et al., 2021: SMOS Near-Real-Time soil moisture processor version 200 for data assimilation at ECMWF. Neural network design and first evaluation results, CESBIO-SMOS report SO-TN-CB-GS-0099 v1.0, 30.06.2021 **[RD7]** Weston, P. and P. de Rosnay: "Quarter 4 2020: Operations Service Report, ESA SMOS ESL contract 4000130567/20/I-BG, January 2021 https://www.ecmwf.int/node/19869 L2 soil moisture operational product v700 read-me-first note can be found here (https://earth.esa.int/eogateway/documents/20142/37627/SMOS-Level-2-Soil-Moisture-release-notes.pdf)

Contact for helpline	For all issues related to data access, formats and read/write, processors, please contact ESA's HelpDesk at eohelp@esa.int.
Comments on the NRT SM Product	Feedback on the NRT soil moisture product can be provided directly to ESA's HelpDesk at eohelp@esa.int.

1. Rationale

The SMOS Near-Real-Time (NRT) soil moisture data (SM-NRT-NN) answer the need of operational agencies to get highly accurate information with minimum data latency. To meet the NRT requirement (i.e. data provision within 3 hours from sensing) the number of auxiliary data files needed in the processor must be kept at a minimum and the processing must be fast. At the same time, the product quality must remain high and be comparable with the product obtained for the operational (geophysical) Level 2 soil moisture (L2) processor, run as part of the ESA ground segment. It is desirable to keep the NRT and operational L2 soil moisture products (L2SM) closely coupled to allow for an efficient product evolution and a common validation procedure.

It was decided to use the SMOS NRT brightness temperature product provided in BUFR format as an input to a Neural Network (NN) that was trained using the operational Level 2 soil moisture product. Version v100 of the SM-NRT-NN dataset was obtained training a NN on version v620 of L2SM data and it has been delivered operationally since early 2016. Version v200 of the SM-NRT-NN dataset was obtained training a NN on version v650 of L2SM data and it has been delivered operationally since August 2018. Essentially, the NN replaces the nominal L2SM processor and NRT brightness temperatures are used instead of the L1C brightness temperatures. Consequently, the two soil moisture data sets, NRT and L2, exhibit the same global climatology but can show local differences. Overall, the NRT product quality is as good as or better than the one of the L2SM data set.

In this note, we briefly describe the processor, the neural network training and the verification of version v300 of the SM-NRT-NN data product. Version v300 of the SM-NRT-NN dataset was obtained training a neural network (NN) on version v700 of L2SM data.

2. Product description and limitations

The Soil Moisture Near Real Time Processor (SM-NRT-OP) provides fast retrieval of soil moisture measurements from the multi-angular brightness temperature as available inside the SMOS level 1 Near Real Time BUFR product. The soil moisture retrieval is based on a neural network scheme trained using the level 2 soil moisture dataset generated by the operational L2 soil moisture processor v700. The input configuration for the Neural Network is a vector of 13 elements containing: six SMOS brightness temperature from 30 to 45 degrees incidence angles in 5 degrees-width bins for both H and V polarization, the six associated local normalization indices reflecting the dynamic range of soil moisture, and the 0-7 cm soil temperature forecast from ECMWF. The neural network architecture has two layers with a hidden layer containing 5 neurons.

The processor consists of two parts: The NRT product generator operated at ECMWF which provides the Soil Moisture NRT product within about four hours from sensing and the offline processor developed and maintained at CESBIO. The offline processor is the essential part for the training of the NN, the quality control, and the definition of the minimum and maximum measurements obtained locally. In the product generation, the NRT brightness temperatures (TBs) are pre-processed and quality controlled, the weights are applied, and the data product is generated.

The data files are then disseminated through EUMETCAST, GTS via ESA and SMOS data portal (https://smos-ds-02.eo.esa.int/oads/access/).

To access the SMOS data see here (https://earth.esa.int/smos/how-to-obtain-data).

The **filename convention** adopted is the following:

Filename convention for the SM-NRT-NN product

Filename:	W_XX-ESA,SMOS,NRTNN_C_LEMM_time1_time2_time3_o_v300_l2sm.nc
Fields description	
W_XX-ESA,SMOS,NRTNN_C_LEMM	Fixed WMO product identifier
time1	Product generation time. Format YYYYMMDDhhmmss
time2	Acquisition time of the first observation into the product. Format YYYYMMDDhhmmss
time3	Acquisition time of the last observation into the product. Format YYYYMMDDhhmmss
o	Fixed one character for operational data.
V300	Processor version used to generate the product
l2sm	Fixed SMOS product identifier
nc	Fixed product format identifier for netCDF
Time format: YYYYMMDDhhmmss, where YYYY corresponds to the year, MM to the month, DD to the day, hh to the hour, mm to the minutes, and ss to the seconds	
Filename example: W_XX-ESA,SMOS,NRTNN_C_LEMM_20160314080722_20160314054231_20160314072114_o_v300_l2sm.nc	

List of parameters available inside the SM-NRT-NN product

SM-NRT-NN NetCDF Field	Units	Description
Latitude	degrees	Geographic latitude of the retrieval
Longitude	degrees	Geographic longitude of the retrieval
Soil moisture	$m^3 m^{-3}$	Soil moisture retrieval value
Soil moisture error	$m^3 m^{-3}$	Estimated uncertainty of the retrieval
RFI probability	%	Probability that the retrieval is affected by RFI. The RFI probability is computed per 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 remaining after filtering.
Number of days since 01-01-2000	days	Acquisition day of the retrieval computed as number of days since 01 January 2001
Number of seconds since midnight UTC	seconds	Acquisition time of the retrieval computed as number of seconds since 00UTC
SMOS DGG id	-	Identification number for the latitude, longitude in the SMOS ISEAS 4H9 grid. The SM-NRT-NN is only provided over land.

The design of the NN and the processing result in some limitations with respect to the operational Level 2 soil moisture product:

- Reduced swath of about 915 Km due to the usage of a limited range of brightness temperature (from 30 to 45 degrees) incidence angle
- Circular gaps in case that brightness temperatures were not measured for all the required incidence angles
- Reduced set of parameters contained in the netCDF structure if compared with the operational Level 2 soil moisture product

Figures 1 and 2 show examples of soil moisture and the associated uncertainties for two SM-NRT-NN v300 files.

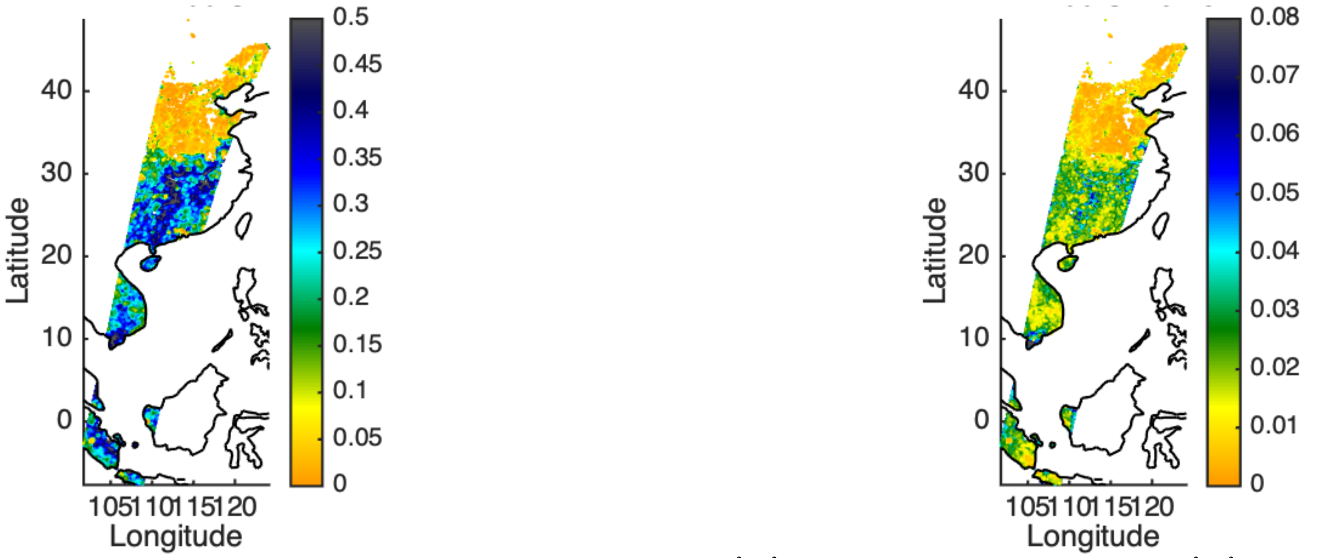

Figure 1: Example of SM-NRT-NN v300 data: soil moisture (left, m^3/m^3) and associated uncertainty (right, m^3/m^3) in file W_XX-ESA,SMOS,NRTNN_C_LEMM_*_ *20190401T102723_20190401T112036*_o_v300_l2sm.nc

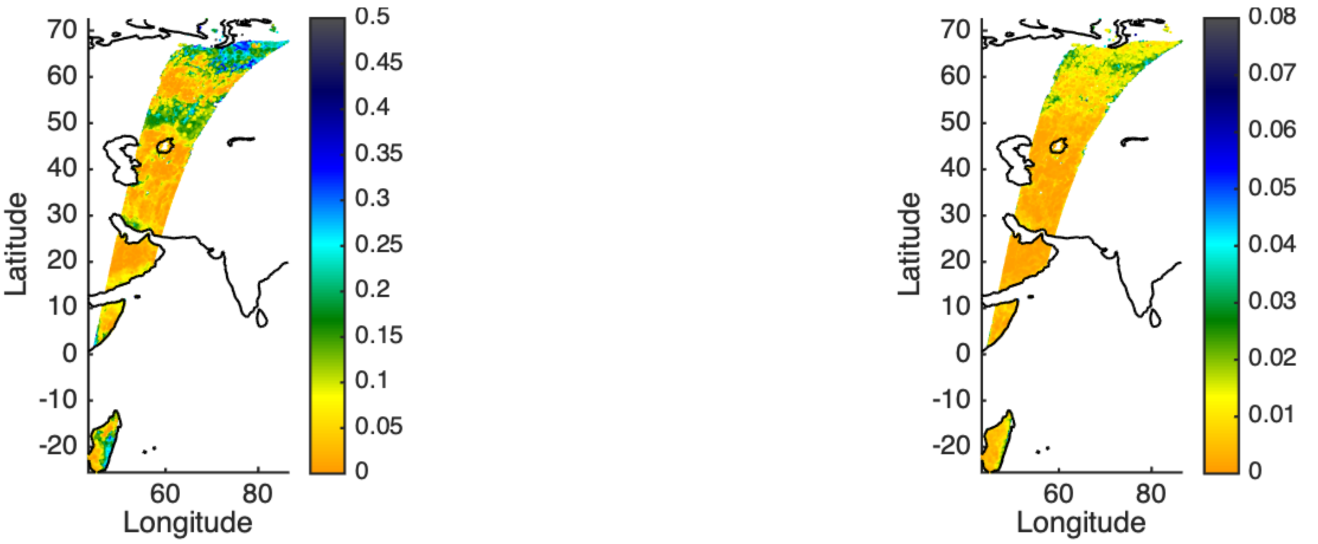

Figure 2: Example of SM-NRT-NN v300 data: soil moisture (left, m^3/m^3) and associated uncertainty (right, m^3/m^3) in file W_XX-ESA,SMOS,NRTNN_C_LEMM_*_ *20191015T141926_20191015T151238*_o_v300_l2sm.nc

3. Training of the Neural Network (NN) for the generation of the SM-NRT-NN product

The NRT brightness temperatures in BUFR format for 2015, July 2016 - June 2017 and April 2019- March 2020 were used for the training and evaluation. Within the first two periods only one day every five was present in the data base. In the third period, all days were present. This data base had also been used to implement v300 of the NRT processing chain specific for data assimilation at ECMWF **[RD6, RD7]**. The input elements, the NN architecture and the training method are described in document **[RD1]** and follow those used to implement v100 and v200, described in documents **[RD2-RD5].** A summary is discussed in the following. The SM-NRT-OP processor has been used to compute a training data base with brightness temperatures in H and V polarizations and three incidence angle bins from 30° to 45°. This configuration is the best trade-off of retrieval accuracy and swath width **[RD2].** The NRT-NN SM is retrieved in swaths of about 915 km (the standard L2 algorithm retrieves soil moisture in swaths of about 1150 km).
The data base (April 2019 - March 2020, and one every five days within the periods Jan 2015 - Dec 2015 and July 2016 - June 2017) was divided in different subsets for the training extracting one day every 20 starting on April 1st, 2019 in the April 2019 - March 2020 period and one day every day 2 of those available in the 2015-2017 period, since in this period only one day every five were present, the actual sampling used for the training data base in one day every 10. Since the database contains more than 5 million samples, two trainings were done using 1/20th and 1/40th of the total number of samples. A subset of 60% of those is used for the actual training, 20% is used for evaluation of the NN performances during the training and to avoid over-training, and the final 20% is used to test the performances of the trained NN a posteriori. Gradient back-propagation and minimization with the Levemberg-Marquardt algorithm has been used. One single hidden layer with 5 neurons has been used, as it has been shown that it is enough to capture the relationship between the input data and the reference SM while keeping 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.

4. Validation of SM-NRT-NN product

Figure 3 shows an example of a retrieval over a few orbits of day 16-May-2017 with L2SM v700 and SM-NRT-NN v300. The performances of the SM-NRT-NN product, generated by the version v300 of the Soil Moisture Near Real Time Processor (SM-NRT-OP), have also been evaluated quantitatively against the operational L2SM v700 product and against in situ measurements.

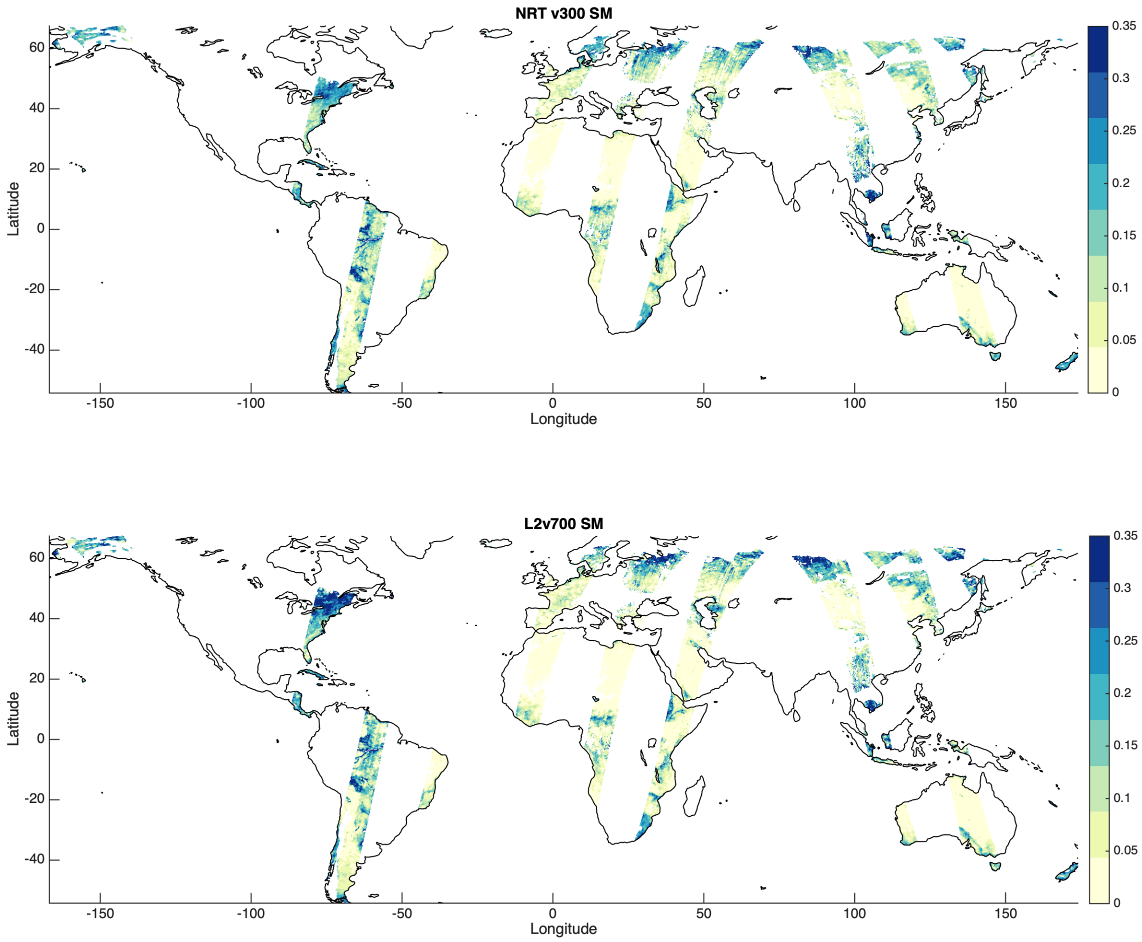

Figure 3: Comparison between SM-NRT-NN v300 (upper panel) and SMOS L2SM v700 (lower panel) for some orbits of 16-May-2017

4.1 SM-NRT-NN v300 comparison to level 2 soil moisture operational product v700

The NRT SM processor has been applied to NRT TBs and compared to SMOS L2SM obtained with version 700 of the SMOS operational processor for all the available data in 2015, 2016, 2017, 2019 and 2020 (Section 3) . Figures 4 and 5 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 L2SM in that period depends on the latitude and the Radio Frequency Interference probability (masked data) but it is higher than 100 (upper panel of Fig. 4) for most of the DGG points.

The bias map in Figure 4 shows that the SM-NRT-NN product has a tendency to underestimate the highest portion of L2SM dataset, which is an 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 room mean square of the differences (RMSD) with respect to the standard deviation of the differences (STDD) in parts of Europe and Canada (Figure 4). However, one should note that both the RMSD and the STDD are lower than 0.04 m3/m3 over most of the land (all except the reddish regions in the two lower panels of Figure 4, see also the positions of the peak of the histograms).

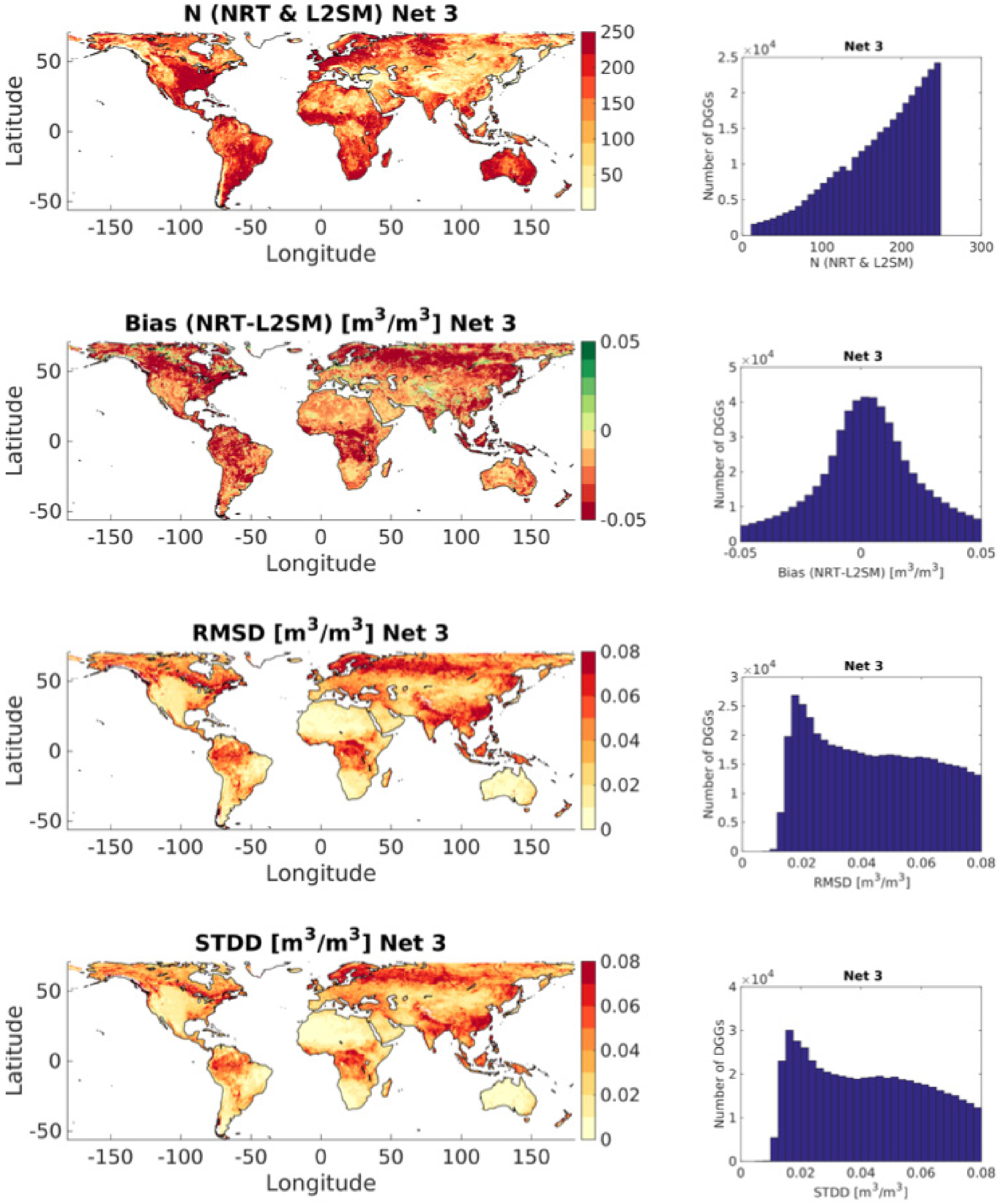

Figure 4: Maps (left) and histograms (right) of (from top to bottom): (i) the number of points (with both SM-NRT-NN and SM L2); (ii) Bias of the SM-NRT-NN product with respect to the L2SM product: mean of SM-NRT-NN minus mean L2SM (m^3/m^3); (iii) Root mean square of the difference (RMSD) (m^3/m^3); (iv) Standard deviation of the difference (STDD) (m^3/m^3).

The correlation of both products shown in Figure 5 is high (> 0.6) 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 is significantly lower over forest (both tropical and boreal). In conclusion, both products show similar dynamics over large parts of the Globe.

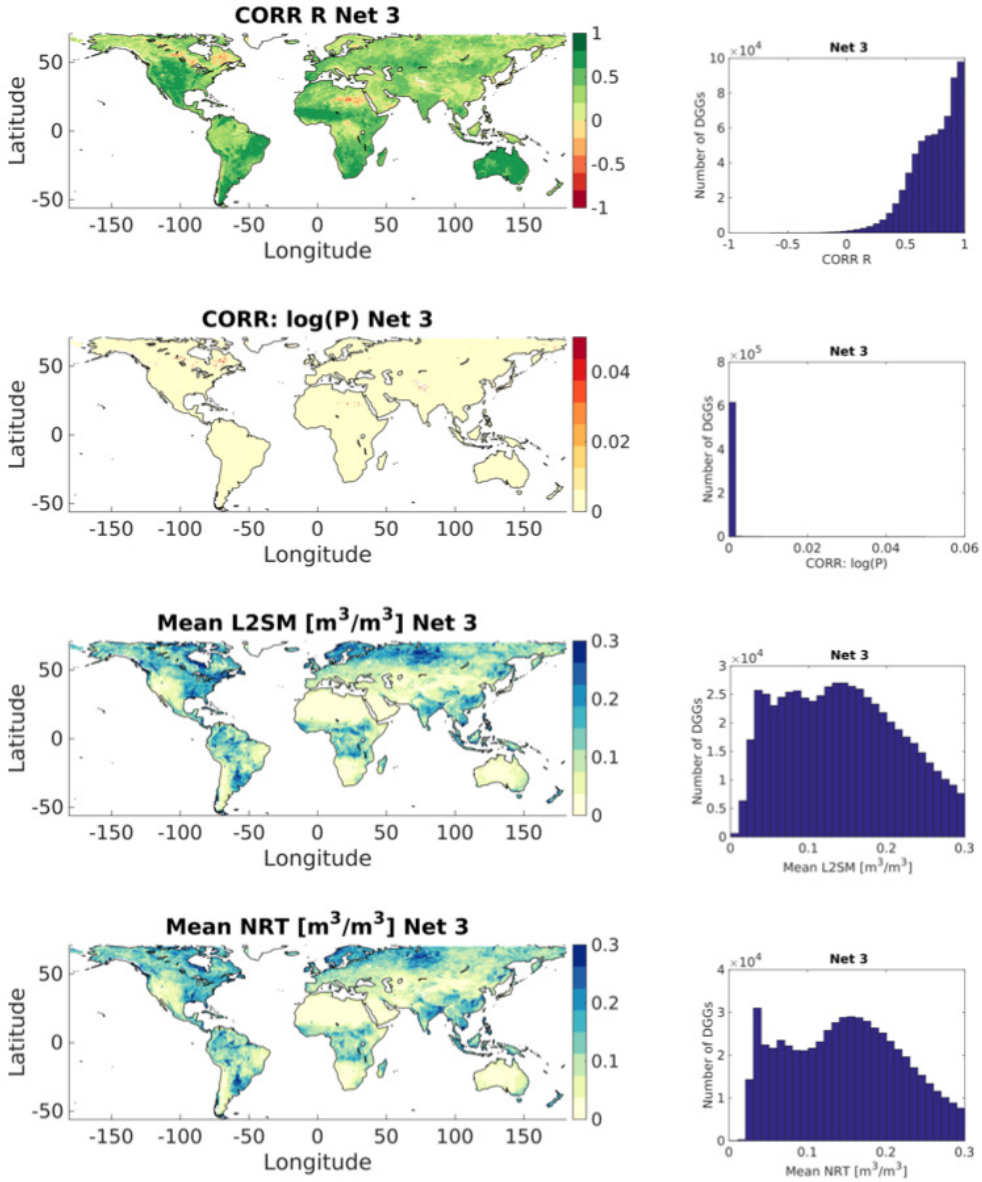

Figure 5: Maps (left) and histograms (right) of (from top to bottom): (i) Pearson correlation (dimensionless) in between SM-NRT-NN and L2SM; (ii) P-value of the Pearson correlation; (iii) Local mean value of L2SM (m^3/m^3); (iv) Local mean value of SM-NRT-NN (m^3/m^3).

Figure 5 also shows the mean of the SM-NRT-NN and L2SM 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 (Figure 4). 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 L2SM dataset.

4.2. SM-NRT-NN comparison to *in situ* measurements

The SM-NRT-NN product has been evaluated against *in situ* measurements from the SCAN, USCRN, SNOTEL, PBO-H2O, REMEDHUS, HOBE, SMOSMANIA, SOILSCAPE, FMI, OZNET, AMMA-CATCH, iRON and TERENO networks (see Rodriguez-Fernandez et al. **[RD1]** and references therein). These networks of *in situ* measurements have been extensively used for the validation of remote sensing data.

Figure 6 shows examples of time series for some sites. 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 resolution 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 L2SM product seems to be closer to the *in situ* measurement and sites for which the NRT product is closer.

Several quality metrics have been computed site per site independently for the SM-NRT-NN and the L2SM products. To show the overall distribution of the quality metrics, boxplots were computed network by network. The results are shown in Figure 7. Both SMOS products show a very similar bias, STDD and R with respect to the *in situ* measurements **[RD1].**

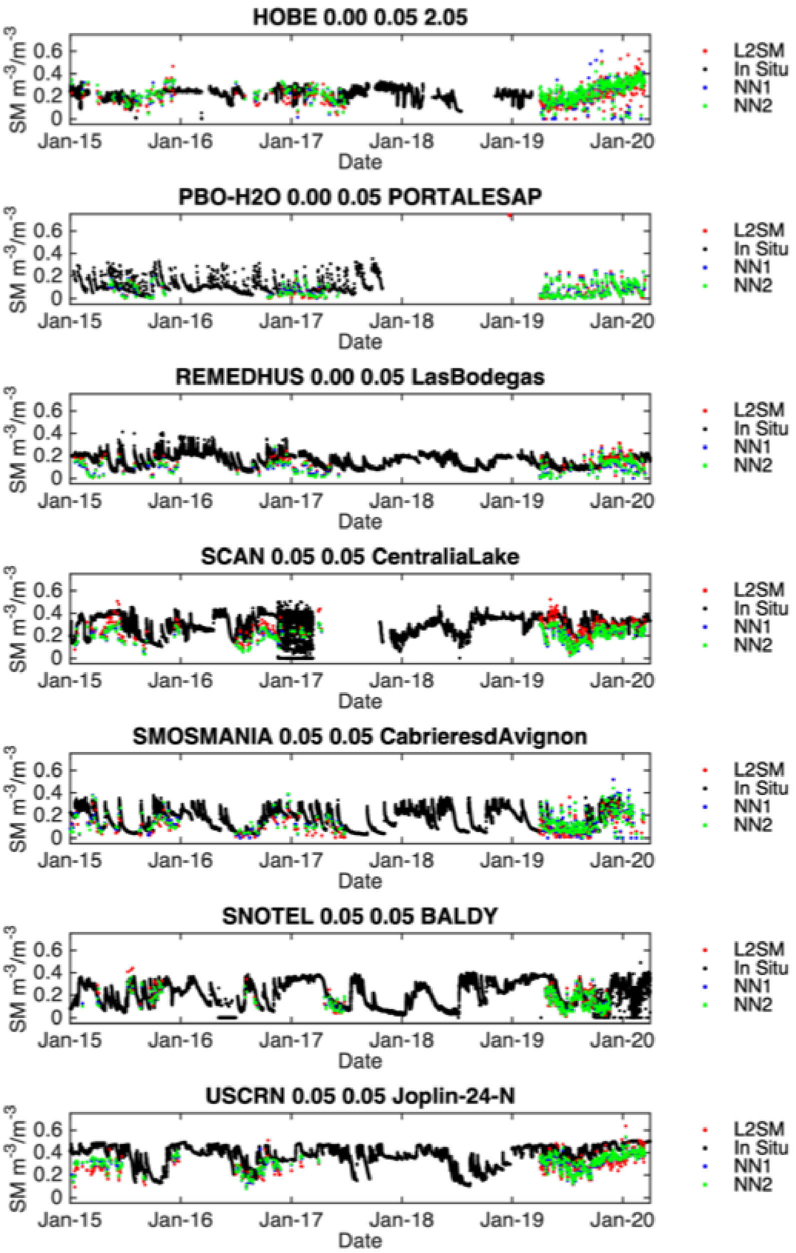

Figure 6: Soil moisture time series for a selection of in situ measurement sites. Black dot: in situ measurements. Red dots: L2SM v700. Green dots: NRT-SM v300 (NN2 of ***[RD1]****). Bleu dots is another dataset (NN1) evaluated in* ***[RD1].***

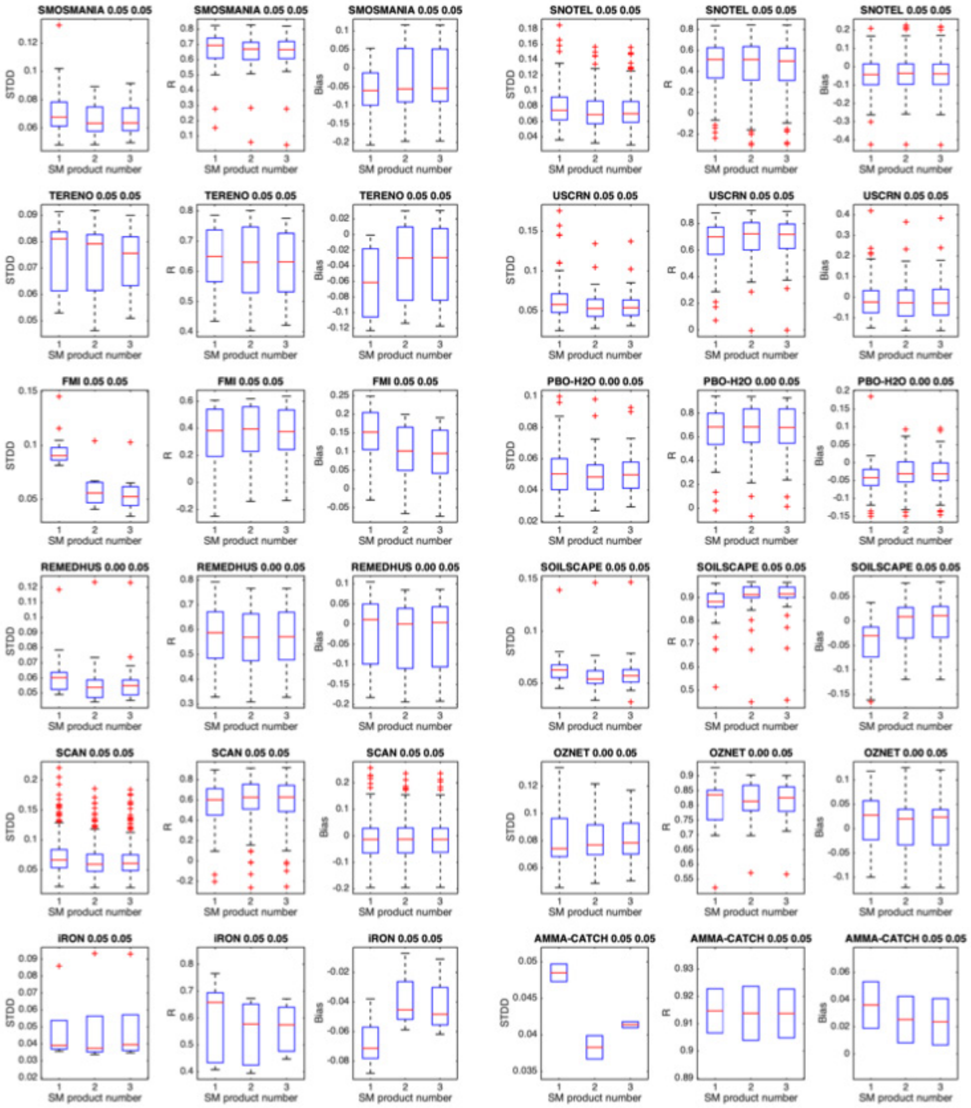

Figure 7: Boxplots for the Pearson correlation coefficient (R), Bias (mean in situ minus mean SMOS SM), and STDD (or unbiased RMSD - ubRMSD), for the L2SM v700 (product 1), NRT v300 candidate NN1 (product 2) and the NRT v300 product (NN2, product 3) in comparison to in situ measurements. The central bar represents the median value of the distribution. The upper edge of the box indicates the 75th percentile of the data set (q_3), and the lower edge indicates the 25th percentile (q_1). 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 $q3 + 1.5(q3 - q1)$ or smaller than $q1 - 1.5(q3 - q1)$.