SAC-D/AQUARIUS SOIL MOISTURE PRODUCT DEVELOPMENT AND EVALUATION FOR PAMPAS PLAINS (ARGENTINA)

Cintia Bruscantini, Francisco Grings, Federico Carballo, Matias Barber, Pablo Perna, Haydee Karszenbaum.*

Instituto de Astronomía y Física del Espacio (IAFE-CONICET-UBA). Buenos Aires, Argentina.
* e-mail: cintiab@iafe.uba.ar

ABSTRACT

In this work, several retrieval algorithms were implemented to retrieve soil moisture (sm) and optical depth (τ) from Aquarius/SAC-D observations. Currently used sm retrieval algorithms (H- and V-pol Single Channel Algorithm, Microwave Polarization Difference Algorithm) were computed over Pampas Plains, Argentina. The methodology of a novel Bayesian algorithm developed was also presented, and its results were contrasted with the previous algorithms. Furthermore, an Artificial Neural Network (ANN) approach to retrieve sm from Aquarius brightness temperature was implemented and trained using SMOS Level-2 sm product. Finally, performance metrics for each algorithm were derived using SMOS L2 sm as benchmark product.

Index Terms— Aquarius; soil moisture; Bayesian inference; Markov Chain Monte Carlo; Artificial Neural Network.

1. INTRODUCTION

Several retrieval algorithms were developed to retrieve soil moisture (sm) from passive remote sensing data. The most commonly used are the Single Channel Algorithm (SCA), the Dual Channel Algorithm (DCA) and Microwave Polarization Difference Algorithm (MPDA). All these algorithms rely on the omega-tao model to link brightness temperature (Tb) and surface dielectric and geometric properties, and differ among them on the polarization channels they use and the minimization scheme implemented [1]. MPDA and DCA make use of H- and V-pol Tb (TbH and TbV) to retrieve sm and optical depth (τ). One disadvantage of both previous algorithms is their sensitivity to noise (mainly uncorrelated noise) in both TbH and TbV. On the other hand, SCAH (SCAV) uses only TbH (TbV) to retrieve sm using τ as an auxiliary input to the retrieval algorithm (usually derived from an optical proxy). The main disadvantage of relying on τ to retrieve sm is that if optical depth is not well known, SCA will have poor performance. In practice, accurate knowledge of τ is tricky. In general, τ is obtained through the vegetation parameter b (a land cover dependent parameter, empirically derived, not unique values found on literature) and vegetation water content, VWC (derived from different proxies and models that result in different VWC values). All these retrieval implementations also need ancillary parameters as necessary auxiliary inputs.

In this work, a novel retrieval algorithm (BRA, Bayesian Retrieval Algorithm) is developed, which uses Bayesian inference to retrieve sm and τ from both H & V channels. The advantages of BRA include: (i) errors on the retrieved variables can be estimated in an univocal way, (ii) prior information about the retrieved variables (provided by other sensors or in situ historical data) can be directly included as inputs to BRA to improve the retrieval, (iii) it can handle uncertainties on the ancillary parameters.

The BRA algorithm uses as a forward model a physical model, zero order radiative transfer (RT-0), that predicts Tb giving a value of sm and ancillary parameters. Another approach considered in this analysis uses an Artificial Neural Network (ANN) to retrieve sm by estimating statistically the link between sm and Tb given a training dataset. Target output dataset was derived from SMOS L2 sm and used to train the ANN.

2. METHODOLOGY

Aquarius/SAC-D sm products were developed using SCAH, SCAV, DCA, MPDA, BRA and ANN algorithms. Soil moisture products were retrieved for Argentina's Pampas Region. The Argentina's Pampas region is located in the center-east of Argentina where the main agricultural activities are cereal production and cattle-raising. It extends over 60 million hectares and accounts for more than 90% of the national grain production. Soybean, wheat, maize and sunflower are the main crops. Weather is among the most important and uncontrollable elements affecting agriculture in this region. Ancillary data used for the retrievals is specific for the area (local land cover and soil texture map). Vegetation optical depth (VWC) was derived from MODIS NDVI [2]. Observations of the MWR 36.5 GHz V-pol channel, Argentinean radiometer on board the SAC-D, was used as proxy of skin temperature over vegetated areas.

This work was funded by MinCyT-CONAE-CONICET project 12.

2.1. Bayesian Inference for solving Soil Moisture Retrieval

The BRA aims to estimate the posterior probability $P_Z(s\bar{m}, \bar{\tau} | TbH_m, TbV_m, \tilde{\theta})$, that is the probability of having mean ground $sm = s\bar{m}$ and $\tau = \bar{\tau}$, given that the sensor (Aquarius) measured TbH_m and TbV_m and that the land ancillary parameters are $\tilde{\theta}$. Estimation of the posterior probability is performed through the Bayes' theorem:

$$P_Z(s\bar{m}, \bar{\tau} | TbH_m, TbV_m, \tilde{\theta}) = \frac{P_L(TbH_m, TbV_m | s\bar{m}, \bar{\tau}, \tilde{\theta}) P_P(s\bar{m}, \bar{\tau})}{\int\int P_L(TbH_m, TbV_m | s\bar{m}, \bar{\tau}, \tilde{\theta}) P_P(s\bar{m}, \bar{\tau}) dsm\, d\tau} \quad (1)$$

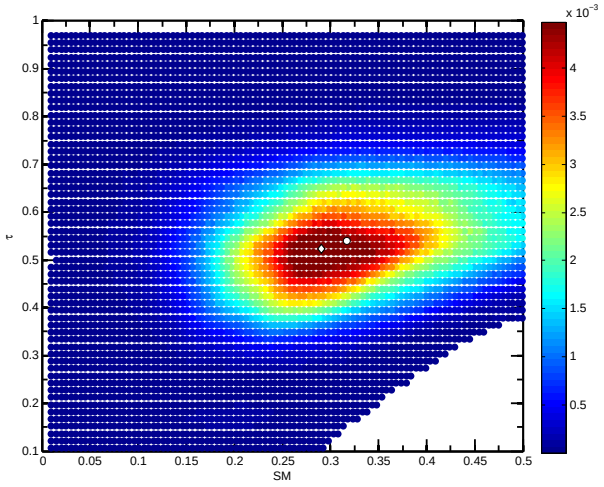
where $P_L(TbH_m, TbV_m | s\bar{m}, \bar{\tau}, \tilde{\theta})$ is the likelihood probability and $P_P(s\bar{m}, \bar{\tau})$ is the prior probability. The likelihood is the probability of the sensor measuring TbH_m and TbV_m when the land conditions are $s\bar{m}, \bar{\tau}$ and $\tilde{\theta}$ (represents the forward model, RT-0 adopted here). If no errors on the ancillary parameters or the forward model are considered, nor instrumental noise on Tb, then the problem is deterministic and the two dimensional likelihood is a delta function centered on the value of sm and τ predicted by the forward model given TbH_m and TbV_m. However, this simplistic assumption is inaccurate. In this work, likelihood is derived in a non parametric manner, in such a way to be a function of ancillary parameters uncertainties (uncertainties in the parameters needed for the retrieval) and instrumental noise. The prior probability is the a priori (before estimation) probability of the variables to be retrieved (sm and τ). Uniform density function would mean that no previous knowledge of the variables was available. On the other hand, a delta function would mean exact previous knowledge of the variables and the BRA would estimate those values independently of the likelihood function. In this work, uniform probability density function (pdf) was considered for sm across the whole possible ranges of values (0 to $0.5\, m^3/m^3$ adopted here) and a Gaussian pdf for τ was adopted, centered on $\tau = b * VWC$, where b is a land cover dependent parameter, and VWC is derived from MODIS NDVI. Gaussian standard deviation is related to accuracy of the τ model ($\tau = b * VWC$) and of the parameters b and VWC to obtain τ. A proxy to VWC uncertainty over soybean crops was derived from the misfit between VWC derived from MODIS NDVI and from Aquarius RVI [3].

Finally, sm and τ can be estimated from the posterior pdf. Both mean and maximum a posteriori (MAP) estimators were considered:

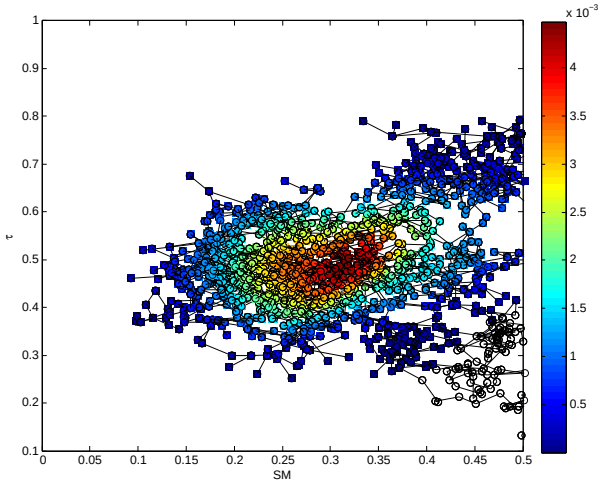
$$\hat{sm}_{mean} = \int\int sm P_Z(s\bar{m}, \bar{\tau} | TbH_m, TbV_m, \tilde{\theta}) dsm\, d\tau \quad (2)$$

$$\hat{sm}_{map} = \arg\max_{sm} P_Z(s\bar{m}, \bar{\tau} | TbH_m, TbV_m, \tilde{\theta}) \quad (3)$$

Furthermore, variance on the retrieved estimations can also be obtained from the posterior pdf for both mean and MAP estimators.



(a) Posterior gridding sampled (◇ MAP, ○ Mean)



(b) Posterior MCMC sampled

Fig. 1. Posterior Sampling with Markov Chain Monte Carlo.

$$\sigma^2_{\hat{sm}_{mean}} = \int\int (sm - \hat{sm}_{mean})^2 P_Z(s\bar{m}, \bar{\tau} | TbH_m, TbV_m, \tilde{\theta}) dsm\, d\tau \quad (4)$$

$$\sigma^2_{\hat{sm}_{map}} = \sigma^2_{\hat{sm}_{mean}} + (\hat{sm}_{mean} - \hat{sm}_{map})^2 \quad (5)$$

The best possible estimator is a fully efficient estimator, which is a minimum variance unbiased (MVU) estimator and will achieve the Cramér-Rao bound (CRB). The Cramér-Rao bound states a lower bound on the variance of the MVU and can be obtained by the following equation:

$$CRB_{sm} = \frac{1}{\mathrm{E}_{\mathrm{sm}}\left[\frac{\partial^2}{\partial sm \partial \tau} \log P_Z(s\bar{m}, \bar{\tau} | TbH_m, TbV_m, \tilde{\theta})\right]} \quad (6)$$

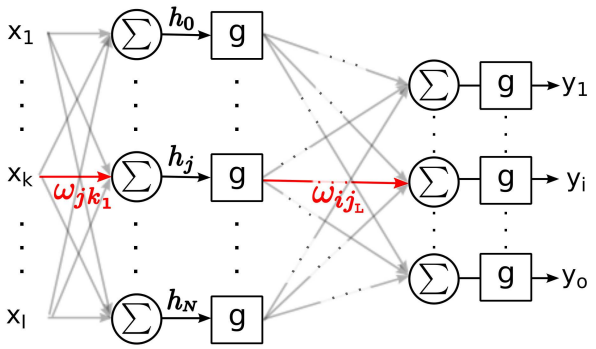


Fig. 2. Neural Network feedforward topology. $\bar{x}$ is the vector of inputs to the network, $\bar{y}$ the outputs vector, $\bar{b}$ the vector of biases, g the transfer function and ω are the synaptic weight matrices. There are I number of inputs, O number of outputs, one output layer and L hidden layers.

2.1.1. Posterior Sampling with Markov Chain Monte Carlo

The main disadvantage of using bayesian inference approach to solve soil moisture retrieval lies in its time performance. Retrieving soil moisture in the area of study for a one week period will take 2 days running in a single core PC. In order to develop an operative bayesian soil moisture algorithm, efforts were made to lower time consumption. A cleverer posterior sampling was carried out implementing Markov Chain Monte Carlo (MCMC) using MPI to parallelise MCMC chains to benefit from multi-core machines or High Performance Computing clusters. The sampling was performed using the Metropolis-Hasting (MH) algorithm. Figure 1 shows an example of posterior sampling on a regular grid (a) and with MH (b). Latter example runs 20 times faster than the regular grid sampling.

2.2. Artificial Neural Network to Retrieve Soil Moisture

A feedforward ANN was implemented to retrieve sm. The topology used for the ANN was a Multi-Layer Perceptron (MLP) such as the one shown in Figure 2, and the learning was performed through Levenberg-Marquard backpropagation algorithm. Several ANN topologies were tested modifying the number of hidden layers (one or two layers) and the number of neurons in each layer.

2.2.1. ANN Inputs & Output Target

Datasets used as inputs to the ANN include: i) Aquarius H- & V-polarization Tb observations of its three beams together with their corresponding incidence angle; ii) MWR 36.5 GHz V-Pol channel was used to estimate canopy temperature [4]; iii) MODIS NDVI was used to obtain $VWC[kg/m^2]$; and iv) Static Parameters: 1) Land-cover-dependent Parameters [5]: ω, b, h; 2) Soil texture [6]: sand, clay.

Dataset used as target output to train the ANN consists on SMOS L2 v5.5.1 sm.

2.2.2. Data sets resolution

Data sets used are from various sources and thus have different temporal and spatial resolution. Therefore, all data sets were averaged (distance weighted) to Aquarius footprints. Only dates were Aquarius, SMOS and MWR data sets were available, were considered for the analysis. MODIS NDVI product has a temporal resolution of 16 days, thus the immediate previous date from Aquarius Tb and SMOS sm was considered. Training dataset period used in the analysis spans from January 1st, 2012 to May 1st, 2013 (excluding August 19th, 2012). Only Aquarius ascending passes and SMOS descending passes were considered (6 pm).

2.2.3. Training Algorithm and Parameters

The dataset was divided into two categories: training and validation datasets. The samples considered for each category were 7000 for training and 3000 for validation of the training. Before training, it is useful to scale the inputs and targets, namely normalization, so that they always fall within a specified range. Without normalization, the input variable with the largest scale will dominate the results. Inputs and targets were scaled so that they fall in the range $[-1, 1]$ by performing a linear transformation on the original data. If the targets were scaled, then the network output will be in the range $[-1, 1]$. In order to convert this output back to its original range, the inverse transformation should be applied.

3. RESULTS

The sm products derived from the BRA approach (Mean and MAP), SCAH, SCAV, MPDA and ANN were computed for August 19th, 2012, over the area of study, and they were evaluated through several performance metrics (correlation, R; bias; root mean square error, RMSE; unbiased RMSE, ubRMSE). SMOS Level-2 sm product was used as benchmark product because, for the date selected, SMOS sm spatial pattern was in good agreement with the Soil Available Water (derived from a water balance model [7]). Nevertheless, absolute SMOS L2 sm values are not necessarily the *ground truth*. Performance metrics results are shown in Table 1.

Metrics for ANN shown in the table are for the 10:10 case (10 neurons in both first and second hidden layers of the ANN), which was proved to have the best performance of all the ANN tested cases (see Figure 3).

4. DISCUSSION

In this work, several retrieval algorithms were implemented for the Pampas Plains, Argentina using mainly Tb observa-

Table 1. Soil Moisture Algorithms Performance Metrics

	R	Bias	RMSE	ubRMSE
BRA Mean	0.811	-0.051	0.094	0.079
BRA Map	0.798	-0.056	0.095	0.077
MPDA	0.728	-0.089	0.123	0.086
SCAH	0.882	0.276	0.366	0.240
SCAV	0.876	-0.051	0.104	0.091
ANN 10:10	0.775	-0.055	0.100	0.083

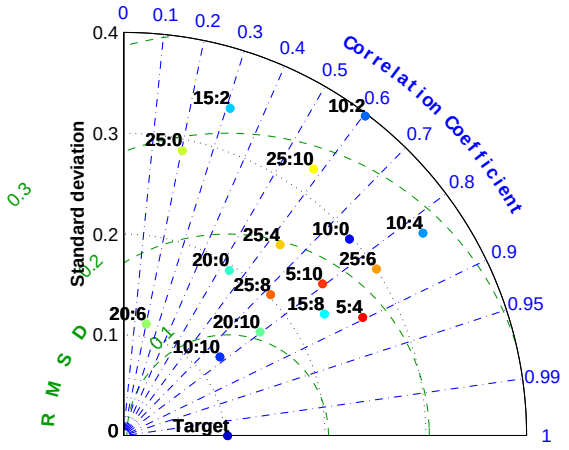


Fig. 3. Performance metrics for different ANN topologies (# neurons on 1st layer : # neurons on 2nd layer).

tions from Aquarius and ancillary parameters from different sensors. All the algorithms make use of almost the same inputs and differ among them on the way that information is used (minimization method, theoretical/statistical model, deterministic/random variables). A Bayesian approach was introduced and an Artificial Neural Network was proposed and trainned. Performance metrics were derived using SMOS L2 sm as a benchmark product.

Results obtained showed that the Bayesian sm products (Mean & MAP) were the ones that showed the lowest ubRMSE, whereas the SCA displayed the highest (in particular H-pol SCA had the poorest performance). Though the ANN was trained with SMOS sm from a different period than the one under test, it was not able to reach the ubRMSE value of BRA. The main difference between ANN and BRA is that, whereas ANN is capable of generalize a training dataset by performing a type of non linear interpolation to find the best functional fit of the inputs and target output, the BRA can manage uncertainties on the parameters taking into account the structure that those errors produce given a theoretical forward model. Thus, the ANN would work better when a vast (spanning the variables domain) and high quality training set is available. On the other hand, the BRA requires a validated forward model and knowledge of the expected errors on the variables input to the model.

MPDA and SCA would run almost as fast as the ANN (making this options eligible for globally operative retrievals products) and their outputs rely on the theoretical forward model. Nevertheless, prior information of sm can not be handled by SCA nor MPDA, as well as errors on the ancillary parameters, as BRA does.

Finally, of all the algorithms implemented, BRA is the only one that, besides retrieving sm and τ, can also provide variance on the retrieved estimations, which might be useful for setting flags and quality control of the product. As a final remark, results point to conclude that the BRA approach is the recommended retrieval algorithm that could be used to validate the selected operative algorithm in specific regions.

5. REFERENCES

[1] C.A. Bruscantini, P. Perna, P. Ferrazzoli, F.M. Grings, H. Karszenbaum, and W.T. Crow, "Effect of forward/inverse model asymmetries over retrieved soil moisture assessed with an OSSE for the Aquarius/SAC-D," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2013.

[2] S. Chan, R. Bindlish, R. Hunt, T. Jackson, and J. Kimball, "Vegetation water content, Preliminary, v.1," *SMAP Ancillary Data Report, Science Document no. 047*, January 2013.

[3] Yihyun Kim, T. Jackson, R. Bindlish, Hoonyol Lee, and Sukyoung Hong, "Radar vegetation index for estimating the vegetation water content of rice and soybean," *Geoscience and Remote Sensing Letters, IEEE*, vol. 9, no. 4, pp. 564–568, July 2012.

[4] M. Owe and A. A. Van De Griend, "On the relationship between thermodynamic surface temperature and high-frequency (37 ghz) vertically polarized brightness temperature under semi-arid conditions," *International Journal of Remote Sensing*, vol. 22, no. 17, pp. 3521–3532, 2001.

[5] INTA (Instituto Nacional de Tecnología Agropecuaria), "Land cover of Argentina 2006-2009," `http://geointa.inta.gov.ar/visor/?p=model_lccs3`, 2014.

[6] INTA SAGyP, "Argentinean Soil Taxonomy - Proyecto PNUD ARG/85/019," `http://geointa.inta.gov.ar/visor/?p=model_lccs3`, 2014.

[7] Argentina Facultad de Agronomía, Universidad de Buenos Aires, "Centro de información agroclimática (CIAg)," `http://www.agro.uba.ar/centros/ciag`.