A Bayesian approach for Soil Moisture and Optical Depth Retrieval: evaluation on Aquarius/SAC-D observations

Cintia Bruscantini, *Student Member, IEEE,* Francisco Grings, Matias Barber, Pablo Perna, Haydee Karszenbaum.

***Abstract*—In this paper, the methodology of a novel Bayesian algorithm to retrieve soil moisture (sm) and optical depth (τ) from passive microwave data is introduced. As a major advantage of this approach, prior knowledge for sm and τ can be directly included within the Bayesian inference framework in order to improve the retrieval. In order to test the methodology presented and contrast its results with existing passive-microwave-based sm retrievals, several algorithms were implemented using Aquarius/SAC-D observations. Algorithms computed include: H- and V-pol Single Channel Algorithm, (SCAH and SCAV respectively) and Microwave Polarization Difference Algorithm (MPDA). Currently available L-band sm products were also incorporated to the study: SCAH for Aquarius (developed by the United States Department of Agriculture) and SMOS Level-2 product (European Space Agency). The analysis was carried out over Pampas Plains, Argentina on an specific date in 2012. Global Land Data Assimilation System sm product was used as benchmark for performance analysis. The Bayesian approach introduced here resulted in the lowest unbiased root mean square error and bias. The main drawback of this approach is that it is highly time consuming, thus making it not suitable for the development of a global near-real time soil moisture product. Efforts were made towards lowering time consumption through Markov Chain Monte Carlo method, although it is still a limiting factor. Nevertheless, the proposed algorithm can provide a framework for evaluation of sm products over limited areas or short time periods.**

***Index Terms*—Aquarius; soil moisture; Bayesian inference; Markov Chain Monte Carlo.**

I. Introduction

Soil moisture (sm) controls the partitioning of net radiation into latent and sensible heat fluxes and the rainfall into infiltration and run-off. Moreover, sm estimation is crucial to monitor water cycle, forecast weather and climate change and to assess flooding and droughts.

Passive microwave data can be used to estimate sm at regional scale for agro-meteorological applications. In the past, several retrieval algorithms were developed to retrieve sm from passive microwave data. Among the most commonly used are the Single Channel Algorithm (SCA) [**?**], the Dual Channel Algorithm (DCA) and the Land Parameter Retrieval Model (LPRM) [**?**]. All these algorithms rely in principle on the same simple physical model, the zero order radiative transfer model [**?**], RT-0, (also called $\omega - \tau$ model), to link the observed brightness temperature (Tb) with surface dielectric and geometric properties (dielectric properties are related to sm). The RT-0 model assumes that observed brightness temperature is an integrated signal coming from the atmosphere, the vegetation and the soil. In particular, the model considers the vegetation as a canopy layer that emits and attenuates microwave signals. In general, at L-band atmospheric contribution is considered negligible.

The main differences between the models considered in this paper are: (1) polarization channels, (2) ancillary data, (3) model parameters, and (4) model assuptions. LPRM and DCA make use of H-pol Tb (TbH) and V-pol Tb (TbV) to retrieve sm and optical depth (τ). One disadvantage of these two algorithms is their sensitivity to noise in both TbH and TbV channels (specially uncorrelated noise between channels). 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 as an external input to retrieve sm is that if τ is not well known, SCA will have poor performance. In practice, accurate knowledge of τ is tricky. In general, τ is modeled as $\tau = b * VWC$, where b is a vegetation parameter (a land cover dependent parameter, empirically derived, not unique values found on literature [**?**]) and VWC is the vegetation water content $[kg/m^2]$(derived from different proxies and models that result in different VWC values [**?**]).

More important, one point in common that all these retrieval implementations share is the need of other biophysical ancillary parameters as necessary auxiliary inputs (e.g. soil and vegetation thermodynamic temperature, soil texture, others). However, spatially distributed information about these ancillary parameters are not known a priori and need to be estimated (in general, from other remotely sensed data). This estimation process are all characterized by non-zero errors. Of course, these errors will propagate through the RT-0 model leading to structural errors on the retrieved variables that are not related to errors in the microwave observations but on the ancillary parameters. In general, these type of errors are discussed but not taken into account on the retrievals based on the above mentioned algorithms.

One systematic way to address these issues is adopting a probabilistic approach that assumes that model parameters are random variables instead of deterministic ones. This implies that although there is a true mean value of a given parameter for a given pixel, the uncertainties related to the estimation of this value can be modeled assuming that the parameter is a random variable. For example, there is a mean soil

C. Bruscantini, F. Grings, M.Barber, P. Perna, and H. Karszenbaum are with the Instituto de Astronoma y Fsica del Espacio (IAFE, CONICET-UBA), Argentina (e-mail: cintiab@iafe.uba.ar).

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

thermodynamic temperature of a 100x100 km pixel (Aquarius footprint resolution), but since the estimation process has nonzero errors, we will assume that the mean soil thermodynamic temperature is a random variable with a variance related to the estimation procedure errors.

In this framework, a novel retrieval algorithm (BRA, Bayesian Retrieval Algorithm) is proposed, which uses Bayesian inference to retrieve sm and τ from both H & V-pol Tb observations. The RT-0 model is used as a forward model to derive the Bayesian likelihood probability density function. Likelihood is derived in a non parametric manner, in such a way to be a function of uncertainties of the parameters needed for the retrieval.

As a major advantage of the Bayesian approach, prior knowledge for sm and τ (e.g. obtained from historical data) could be directly included as inputs to BRA to improve the retrieval. In this paper, NDVI (Normalized Difference Vegetation Index) and RVI (Radar Vegetation Index) information were used to derive a pixel-by-pixel prior of VWC although a non-informative prior distribution was used for sm.

This work is organized as follows. First, BRA approach is presented and its pros and cons are discussed and addressed. Second, a first version of BRA algorithm is implemented using the L-band Tb observations of the Aquarius radiometer on board the SAC-D platform over a region in Argentina. Results are contrasted with other existing sm algorithms, some of them currently available (such as global sm products developed by the United States Department of Agriculture [**?**] for Aquarius and by the European Space Agency for SMOS mission [**?**]) and some of them were implemented ad hoc for completeness (H- and V-pol SCA, SCAH, SCAV; Microwave Polarization Difference Algorithm, MPDA). Finally, a performance analysis using the Global Land Data Assimilation System as benchmark is presented and discussed.

II. Methodology

A. Bayesian Algorithm

Current sm retrieval algorithms such as MPDA and SCA make use of the RT-0 model in a deterministic way, minimizing the difference between observed and predicted values in order to retrieve sm (and τ in the case of MPDA). These models do not have a way to model the uncertainty of the instruments, model coefficients, and ancillary data. In this paper, we will include in the analysis model parameters uncertainties and instrumental noise.

Bayes' theorem, which can be derived from basic conditional probability rules, states that,

$$P(A|B,I) = \frac{P(B|A,I)P(A|I)}{P(B|I)} \tag{1}$$

where A, B and I are propositions and $P(A|B,I)$ is the probability of proposition A given proposition B and I. So stated, this powerful theorem allows to compute the probability of the variables to be inferred (sm and τ in this case) as a function of the observed variables (TbH and TbV) and auxiliary information (ancillary parameters) if we know a functional relation between the observed and inferred variables. This functional relation $[TbH, TbV] = f(sm, \tau, \bar{I})$ is of course the forward model. For historical reasons, the terms in Eq. (1) are given specific names: $P(A|B,I)$ is the *posterior* distribution of A given B and I, $P(B|A,I)$ is the *likelihood* of B given A and I, $P(A|I)$ is the *prior* distribution of A and $P(B|I)$ is the *evidence*.

This formalism can be directly applied to the estimation of sm and τ given TbH and TbV as follows. Using Bayes' theorem, the conditional (*posterior*) probability of having the terrain condition $s\bar{m}$ and $\bar{\tau}$ given measured TbH and TbV (TbH_m and TbV_m) and the ancillary parameters $\theta = \tilde{\theta}$ can be expressed as follows:

$$P_Z(s\bar{m}, \bar{\tau}|TbH_m, TbV_m, \tilde{\theta}) = \frac{P_L(TbH_m, TbV_m|s\bar{m}, \bar{\tau}, \tilde{\theta})P_P(s\bar{m}, \bar{\tau})}{\int\int_D P_L(TbH_m, TbV_m|s\bar{m}, \bar{\tau}, \tilde{\theta})P_P(s\bar{m}, \bar{\tau})dsm\, d\tau} \tag{2}$$

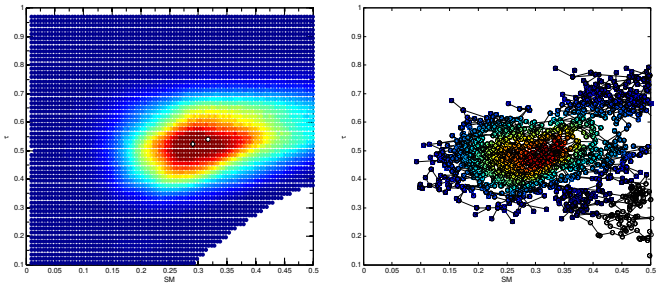
being D the sm and τ domain in which the forward model is valid, $P_L(TbH_m, TbV_m|s\bar{m}, \bar{\tau}, \tilde{\theta})$ is the (*likelihood*) probability of measuring TbH_m and TbV_m given the terrain state $sm = s\bar{m}$, $\tau = \bar{\tau}$ and $\theta = \tilde{\theta}$, $P_P(s\bar{m}, \bar{\tau})$ is the (*prior*) joint density function of $s\bar{m}$ and $\bar{\tau}$ (that includes previous knowledge of $s\bar{m}$ and $\bar{\tau}$), and the double integral is a normalization factor (*evidence*) that computes the overall probability of measuring TbH_m and TbV_m.

1) Bayesian Algorithm: Likelihood construction and computation: The *likelihood* is a function of both the pair [sm, τ] (the variables we want to estimate) and the ancillary parameters θ (obtained from auxiliary information). Therefore, the *likelihood* is related to both the forward model structure (in this case, RT-0) and the distribution of model's parameters values. If no uncertainties are considered on measured Tb nor in ancillary parameters, then the *likelihood* becomes a delta function at [sm, τ] that corresponds to the estimation of RT-0 given measured TbH and TbV for a given θ. As the uncertainties on θ and measured Tb increase, the *likelihood* spreads in the [sm, τ] space in a particular manner, following the structure of the model function and the distribution of the model's parameters. In the case of study presented in Section III, normal distributions were assumed for the ancillary parameters, where mean values are the ones provided by the auxiliary information and the variances are related to educated guesses of parameter errors (see Table I). Instrumental noise was also considered by adding noise to the TbH-TbV pairs computed by the RT-0 in these conditions (Gaussian distributed with zero mean and 0.5K standard deviation, typical instrument error).

Once defined the pdfs, the computation of the *likelihood* pdf is performed as follows. First, samples of the θ random variables are obtained, second, several pairs of TbH-TbV are predicted by RT-0 for a fixed pair of sm-τ values (sm_i and τ_i). Third, using a non parametric method (kernel smoother), a pdf is derived from the 2-dimensional scatter plot of TbH and TbV. Finally, the pdf is evaluated at TbH_m and TbV_m (measured values of Tb), that accounts for the probability of sm_i and τ_i being equal to the ground truth given that Tb observations were TbH_m and TbV_m. This procedure evaluates the *likelihood* at sm_i and τ_i, and it should be repeated as many

Parameter	Uncertainty	Source of Uncertainty
$Sand$	15%	Soil texture triangle [**?**]
$Clay$	10%	Soil texture triangle [**?**]
Tbh/Tbv	0.5K	NEDT worst case scenario [**?**]
$Tsoil$	4K	Residues of MWR vs. Windsat cross-calibration [**?**]
ω	0.05	Land-cover dependent Look Up Table [**?**] [**?**]
h	0.02	Regression residues [**?**]

TABLE I: Parameter uncertainties considered for BRA approach.



(a) (b)

Fig. 1. Two different methodologies of posterior sampling: regular grid (◇ MAP, ○ Mean) (a) and Markov Chain Monte Carlo (b). Color indicates probability, being reddish (bluish) higher (lower) probability.

times as points in the *likelihood* grid.

2) Bayesian Algorithm: Prior construction and computation: The *prior* should be defined on all the sm and τ domain and allows us to assign some probability distribution to the retrievable variables before performing the estimation. Thus, any previous knowledge of sm or τ can be included in the retrieval to constrained the estimation. Examples of sources of such previous information can include: land surface models, climatology, past estimation from another systems, field measurements, satellite-based products, etc. The *prior* used in the case of study presented in Section III was chosen to be uniform ranging from 0 to $0.5\,m^3/m^3$ for the sm variable, since no previous knowledge of sm was considered besides its possible range. The *prior* for τ was assumed normal, centered on the τ value derived from MODIS NDVI [**?**], with a variance being a linear relation of the absolute difference between τ obtained from MODIS NDVI (τ_{NDVI}) and Aquarius RVI (τ_{RVI}) [**?**]. When both τ_{NDVI} and τ_{RVI} are similar, then it is assumed that the derived τ value is reliable. Accordingly, the *prior* pdf enhances the *likelihood* pdf on the area of the domain where τ values are close to the τ_{NDVI}. Therefore, in this case, the *prior* pdf highly restricts the *posterior* pdf, thus strongly lowering its variance, consequently lowering the variance on the retrieved variables. On the other hand, if τ_{NDVI} and τ_{RVI} are very different, then the τ_{NDVI} may not be an accurate estimation of τ, and the *posterior* pdf is likely to resemble the *likelihood*.

Finally, if the uncertainties on the ancillary parameters are low, then the BRA approach is presumably to encounter sm and τ values similar to the ones retrieved by the MPDA, and with a variance related to the degree of uncertainty on $\tilde{\theta}$. It should be noticed that relations between VWC and RVI have been derived for soybean and rice land covers [**?**], thus τ_{RVI} could be computed on limited areas. Elsewhere, Gaussian variance was remained fixed to 0.1, a rather loose condition.

3) Bayesian Algorithm: Estimators: Given the *posterior* pdf in (2), two estimators were derived. One is the minimum variance estimator, *expected a posteriori* (BRA Mean). It is derived as the expectation value of the posterior pdf $\hat{sm}_{mean}$, that has variance $\sigma^2_{\hat{sm}_{mean}}$. The other estimator implemented was the *maximum a posteriori* (BRA MAP), which is the mode of the *posterior* pdf, $\hat{sm}_{map}$, with variance $\sigma^2_{\hat{sm}_{map}}$. Both estimators were also used to estimate τ with its corresponding functional form.

The advantages of BRA are: (i) errors on the retrieved variables can be estimated in an unequivocal way, (ii) it gives the possibility to use prior information about the retrieved variables (provided by other sensors or *in situ* historical data), (iii) it can handle uncertainties on the ancillary parameters, (iv) intervals can be retrieved for a given confidence level. The main disadvantage of BRA is its time performance. In order to lower runtime, a Markov Chain Monte Carlo was implemented.

4) Bayesian Algorithm: Markov Chain Monte Carlo for Posterior sampling: In a preliminary version of the algorithm, BRA approach was computed on a regular grid spanning Bayesian pdfs domain (limited mainly by the *prior* pdf). In this scheme, precision of the estimations are related to grid resolution. Therefore, to lower time computation of the algorithm, a coarse grid was used at first, and then it was refined over a subarea of the domain where the *posterior* pdf displayed significant probability values. An even better sampling approach involves the implementation of a Markov Chain Monte Carlo (MCMC) method, which consists in random walks sampling efficiently the probability distributions. Differences between older and newer version samplers are shown in Fig. 1. For MCMC method, Metropolis Hasting was selected as the generator algorithm of the random walks. As a consequence of this newer sampler, MCMC resulted in a 10x speedup using an 8-cores CPU. Proper MCMC was implemented by carefully addressing the following requirements: i) sufficient initial *burn in* iterations and ii) fulfill of convergence criteria.

III. Case of Study

In order to test the proposed methodology and contrast it with existing sm algorithms, BRA approach was implemented using Tb observations of Aquarius radiometer. Launched on June 2011, the Aquarius/SAC-D mission is an international cooperation between CONAE (Comisión Nacional de Actividades Espaciales), Argentina, and NASA, USA. Aquarius is an integrated L-band radiometer (1.413 GHz) and scatterometer (1.26 GHz). Aquarius radiometer has three cross-track beams that scan in a push-broom fashion at incidence angles of 28.7°, 37.8°and 45.6°. Its primary goal is to monitor weekly

global sea surface salinity to help understanding both climate change and the global water cycle [**?**]. However, land Aquarius observations can be used for monitoring sm on a global scale at a rather coarse resolution ($\sim$ 100 km). Thus, in this work, sm retrieval from Aquarius 3-beams Tb observations was computed over a limited area of study. The Pampas Plains region in Argentina was selected due to the fact of being the most important agricultural area of Argentina.

A. Study area and ancillary parameters

The Argentina's Pampas region is a wide plain located in the center-east of Argentina (27-40 °S, 57-67 °W) where main agricultural activities are cereal production and cattle-raising. It extends approximately 50 million hectares of fertile lands and accounts for more than 90% of the national grain production. Main crops include soybean, wheat, maize and sunflower. In particular, soybean extends over a wide area within the region (τ_{RVI} can be locally computed). Weather is among the most important and uncontrollable elements affecting agriculture in this region. Most of the Pampas region is significantly affected by cyclical drought and flood episodes that impact both crop and cattle production. In general, along the region, the area is drier in the west and becomes wetter in the east.

Customized inputs for this area include specific ancillary parameters (land cover [**?**] and soil texture [**?**]). Land cover dependent parameters (ω, single scattering albedo; b, vegetation parameter related to optical depth; h, roughness parameter; stem factor) were selected following the Look Up Table of algorithm parameters on the Soil Moisture Active Passive (SMAP) Algorithm Theoretical Basis Document [**?**]. MODIS MOD13Q1 product provides NDVI every 16 days at 250-meter spatial resolution and it was used to derive high-resolution $VWC\,[kg/m^2]$ following [**?**]. Subsequently, VWC was non linearly aggregated to Aquarius coarse resolution following the methodology in [**?**]. Soil temperature (assumed to be the same as canopy temperature) was derived from the Microwave Radiometer (MWR). The MWR is a three channel Dicke radiometer on board the SAC-D spacecraft and operates at 23.8 GHz H-Pol and 36.5 GHz V- & H-Pol. It provides near-simultaneous and spatially collocated observations with Aquarius measurements. MWR channel 36.5 GHz V-Pol observations were used to estimate soil and canopy temperature following [**?**].

Inputs to the BRA approach also include the uncertainty considered in each random variable (ancillary parameters of the RT-0 model and Aquarius Tb observations). Values considered are listed in Table I.

B. Results

Sm products were derived using the BRA approach (Mean and MAP) over the area of study for two Aquarius and SMOS overpasses on August 2012 (austral winter, low vegetation, overall wet soil conditions). In addition to implementing the BRA algorithm, other algorithms were computed. The algorithms selected for the comparison are SCAH, SCAV and

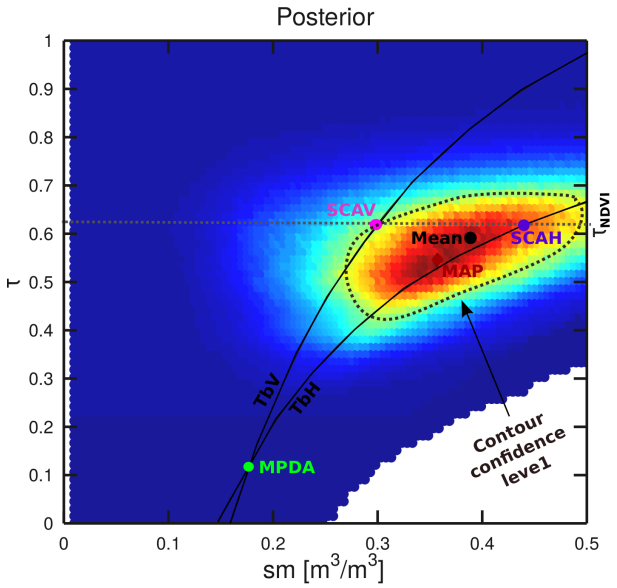


Fig. 2. Interpretation of the retrieval algorithms implemented in this analysis.

MPDA. MPDA is based on LPRM algorithm **??**, though L-band parametrization is used and was selected to match SMAP parameter values defined in the Algorithm Theoretical Basis Documents [**?**]. Ancillary parameters for all algorithms were selected to be consistent.

For a better understanding of the differences between the algorithms, a retrieval example is shown in Fig. 2. The RT-0 model contour level curves (evaluated at ancillary parameters θ_i and Aquarius observations TbH and TbV) are marked as TbH and TbV. The MPDA, that uses both H- and V- channels, retrieves the values of sm and τ where both curves intersect. SCAH (SCAV) uses TbH (TbV) and τ (derived from NDVI) as inputs to retrieve sm. If considered τ value is unpolarized and different from τ value retrieved by MPDA, then different sm values are retrieved from SCAH and SCAV (as shown in the current example). [1]

As mentioned in Section II, BRA approach uses RT-0 to compute the *likelihood* pdf. The *likelihood* provides a probability for each [sm, τ] pair, because the ancillary parameters were considered to be random variables instead of deterministic values. As stated, the *prior* considered is uniform on sm and Gaussian on τ (centered on τ_{NDVI}). The *prior* enhances the *likelihood* distribution close to the $\tau = \tau_{NDVI}$ domain. The posterior distribution is shown in Fig. 2 as the colored background image, where reddish (bluish) colors indicate higher (lower) probability values. Given this *posterior*, the Mean and MAP estimations are computed as the expectation and the mode of the distribution respectively. [2]

[1] Nevertheless, if τ was polarized, then some combinations of τ_H and τ_V can produce the same sm values for both SCA algorithms.

[2] If relevant, knowledge of the *posterior* pdf allows to retrieve sm intervals given a level of confidence (α) on the estimation (using the *posterior* contour curves such as the one shown in Fig. 2). In this case, where MCMC is used, approximating *posterior* contour curves is extremely difficult and requires too many samples, thus equal-tailed intervals method is preferred where intervals are constructed from the $\alpha/2$ to 1-$\alpha/2$ quantiles of the simulated dataset.

		SMOS	MPDA	USDA	SCAH	SCAV
Mean	***r***	0.813	0.697	0.836	0.784	0.770
	bias	0.049	-0.040	0.082	0.319	0.099
	RMSE	0.091	0.075	0.124	0.429	0.162
	ubRMSE	0.076	0.063	0.092	0.287	0.128
MAP	***r***	0.794	0.691	0.789	0.762	0.784
	bias	0.054	-0.034	0.089	0.324	0.104
	RMSE	0.093	0.075	0.131	0.429	0.160
	ubRMSE	0.076	0.067	0.096	0.282	0.121

TABLE II: Soil Moisture Algorithms Performance Metrics

	r	***bias***	***RMSE***	**ubRMSE**
Mean	0.900	0.017	0.034	0.030
MAP	0.890	0.014	0.035	0.032
SMOS	0.921	0.075	0.094	0.058
MPDA	0.756	-0.022	0.057	0.053
USDA	0.902	0.113	0.140	0.082
SCAH	0.845	0.377	0.463	0.268
SCAV	0.859	0.130	0.173	0.115

TABLE III: Performance Metrics using GLDAS sm as benchmark

In order to extend the comparative analysis, Aquarius Level 2 sm product provided by United States Department of Agriculture (USDA) [**?**] and SMOS Level 2 sm product [**?**] were also evaluated over the area of study. It is important to point out the differences on SMOS footprint resolution (40 km nominal) and Aquarius (100 km depending on the beam). Accordingly, SMOS sm was distance-weighted upscale to Aquarius resolution.

All sm maps are shown in Fig. 3. Although there are not *in situ* measurements to validate the sm products, a Soil Available Water (SAW) product (derived from a water balance model [**?**]) was included in the analysis for visual inspection. In general, all sm spatial patterns are in good agreement with this product, but since SAW and sm are different variables, direct quantitative comparison cannot be carried out.

In Fig. 3 is noteworthy the remarkable differences on the sm dynamic range of each algorithm. Whereas SMOS, SCAH and SCAV displayed sm values as high as $0.6\,m^3/m^3$, USDA, MPDA and BRA saturate at sm around $0.5\,m^3/m^3$. USDA algorithm saturates sm *manually* by taking into account the field capacity (around $0.55\,m^3/m^3$ depending on the soil texture), and BRA approaches were *by design* saturated at $0.5\,m^3/m^3$ by assigning zero probability to sm higher than $0.5\,m^3/m^3$ on the *prior* pdf.

A quantitative analysis was carried out to compare BRA and the other algorithms results by means of several performance metrics (correlation, *r*; bias; root mean square error, RMSE; unbiased RMSE, ubRMSE). Computation of such metrics is discussed in [**?**]. Performance metrics results are shown in Table II.

Pearson correlation coefficient (*r*) shows that linear dependence between BRA (Mean and MAP) and USDA, firstly, followed by SMOS, was the most significant. The lowest correlation was found with MPDA and BRA. However, MPDA displayed the lowest ubRMSE. This metrics are useful to compare the new Bayesian approach described here with the already established retrieval algorithms. Nevertheless, no conclusions can be drawn on absolute performance. Thus, in the next section, a sm product derived from a land surface model is introduced in this analysis to serve as benchmark dataset.

1) Using GLDAS as benchmark: Since there are no *in situ* sm data available at the scale of Aquarius in Pampas region, an alternative methodology to evaluate the performance of the algorithms was established. To this end, we introduced in the analysis a well known model which is commonly used to estimated sm at global scale. The Global Land Data Assimilation System (GLDAS) is a global, off-line (uncoupled to the atmosphere) terrestrial modeling system developed by NASA and NOAA that uses both ground and satellite observations as forcing of advanced land surface models, integrated to data assimilation techniques that generates optimal fields of land surface states (soil moisture among them).

In this paper, we used the 0-10 cm depth sm product provided by GLDAS version 1, NOAH Land Surface Model, at 1° resolution grid and 3-hours temporal resolution (the closest to Aquarius overpass time, see Fig. 3i). Accordingly, all previous sm products were interpolated (distance-weighted) to GLDAS grid.

Fig. 4 shows a comparison between L-band (SMOS and Aquarius) retrieved sm values, and GLDAS sm. Performance metrics were also derived for this comparison and are listed on Table III.

In Fig. 4, it can be seen that BRA estimated values are very similar to the ones computed by GLDAS, with only a few points outside the estimation error. On Table III, it can be seen that BRA was the algorithm that showed the lowest bias and ubRMSE. SMOS, USDA and BRA displayed the highest values of *r*. On the other hand, the highest ubRMSE and bias was obtained by SCAH, that displayed too high sm values which are unrealistic (higher than $1\,m^3/m^3$). Finally, MPDA exhibited the lowest *r*.

This overall good performance of BRA over other candidates algorithms studied in this example is not surprising, since a Bayesian approach is a generalization of most of this models.

Vegetation water content (and therefore τ) is the most significant parameter that impacts SCA sm retrieval [**?**]. Poor estimation of τ will result in SCA poor performance. In fact, SCAH (SCAV) algorithm is not sensitive to uncertainties in TbV (TbH), but strongly rely on τ estimations. If τ values used as input on the SCA retrievals are overestimated, retrieved sm will also be overestimated. In this context, the most common approach to correct systematic errors on τ values is to change the b parameter. In effect, USDA SCAH uses lower b values (globally constant $b = 0.08$) than the ones considered in this paper (land-cover dependent b ranging from 0.10 to 0.13 following [**?**]), achieving a better performance. However, USDA sm still displays significantly higher values than GLDAS sm (see Fig. 4b). On the other hand, for the cases of dual channel algorithms (i.e. MPDA), there is no sensitivity

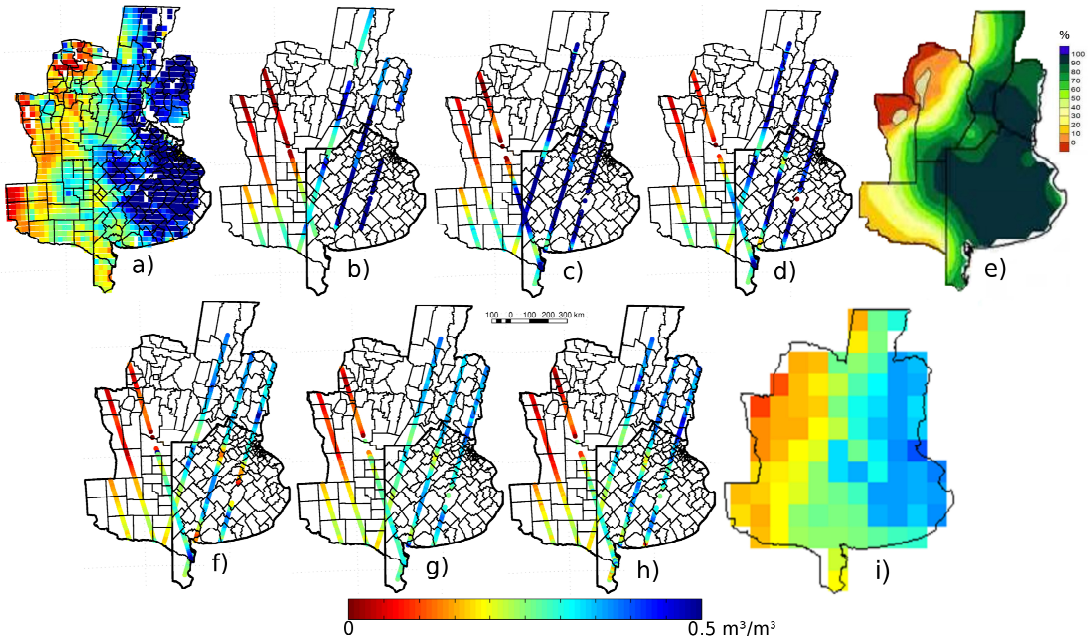


Fig. 3. Sm retrieved by SMOS (a), USDA SCAH (b), SCAH (c), SCAV (d), Soil Available Water (e), MPDA (f), BRA Mean (g), BRA MAP (h), GLDAS (i) over the Pampas Plains region in Argentina. Differences on the geometry of sm acquisition between SMOS and Aquarius can be observed. Figures (b)-(d) and (f)-(h) show Aquarius three beams ascending and descending passes (ascending inner, middle and outer beams and descending outer, middle and inner beams from left to right).

to external estimation of τ (not used), but a known sensitivity to errors on TbH and TbV (specially to uncorrelated errors, uncertainties on the quotient TbH/TbV), since a small change in Tb values could imply a large change in estimated sm.

These two problems are automatically addressed in this Bayesian approach, first by considering a distribution of b values and a prior distribution for τ and second by considering uncertainties on TbH and TbV (related to instrument and/or θ errors). In this way, instead of just forcing the solution to unrealistic sm retrievals when input or ancillary data are dubious, the Bayesian algorithm automatically weights the uncertainties and points to a more conservative retrieval. This retrieval is of course a function of the uncertainties selected for θ parameters. BRA retrievals too similar to MPDA might indicate that the *prior* pdf is not too restrictive and/or that the uncertainties considered on θ were relatively small. Though θ Gaussian pdf are symmetric, RT-0 is a nonlinear model, thus creating a bias between BRA and MPDA sm and τ retrievals. This bias is not an artifact, but a correction in the $[sm, \tau]$ estimations due to uncertainties in θ.

As previously mentioned, one important feature of BRA approach is that it can retrieve quality flags by means of the *posterior* pdf. BRA error bars shown in Fig. 4 represent standard deviation (σ) of retrieved sm. Values of σ are computed as the standard deviation of the samples in MCMC chains. In general, at both BRA Mean and MAP, error bars increase at higher retrieved sm values. This is consistent with the decreased sensitivity of the RT-0 model at high sm values (see RT-0 level curves in Fig. 2).

Finally, it is instructive to briefly discuss the effect of the prior distribution over τ. In general, estimations of VWC from NDVI had significant differences with the ones obtained from Aquarius RVI for soybean land cover. In our approach, this implies a rather loose *prior* pdf. This relatively uninformative pdf does not constrain the retrievals in a determinant way, in accordance to our information about VWC, that indicates that our two VWC proxies strongly disagree on their estimations.

IV. Discussion

A new passive-based sm and τ retrieval algorithm that makes use of Bayesian inference was proposed and its performance was evaluated. As major advantages, the BRA approach provides errors on the estimated variables, enable to enter prior knowledge of the variables to be retrieved and can manage uncertainties on the ancillary parameters.

A case of study was considered to test the BRA algorithm using Aquarius/SAC-D observations over Pampas Plains region in Argentina. Several other sm retrieval algorithms were implemented (SCAH, SCAV, MPDA) and existing sm products (USDA, SMOS) were evaluated. In absence of adequate *in situ* data, results were contrasted using as benchmark GLDAS sm product. Performance metrics for each retrieval algorithm were derived. BRA exhibited the lowest bias, RMSE (and ubRMSE consequently) and higher correlation to GLDAS.

Reasons for BRA good performance were analyzed. In particular, three main features were addressed. First, extreme sm values: SMOS, SCAH and SCAV displayed sm values as high as $0.6\,m^3/m^3$, whereas BRA prior considered no probability for sm higher than $0.5\,m^3/m^3$ and USDA saturate sm values at field capacity (around $0.55\,m^3/m^3$ depending on the soil porosity). Both BRA and MPDA approaches retrieved sm values lower than $0.45\,m^3/m^3$ and GLDAS

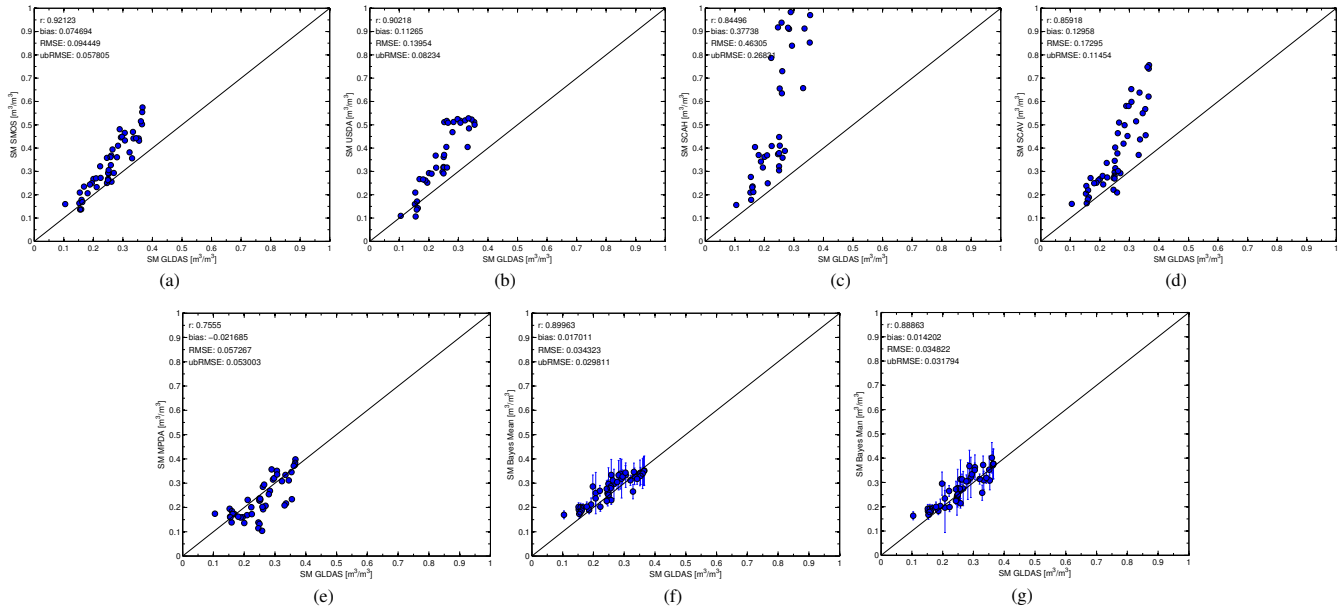


Fig. 4. Comparison of retrieved sm between GLDAS and SMOS (a) USDA SCAH (b) SCAH (c) SCAV (d) MPDA (e) BRA Mean (f) BRA MAP (g) algorithms for the case of study analyzed.

lower than $0.4\,m^3/m^3$. Second, possible mismatches on the parameterizations were also considered, specially on the b parameter and τ estimations from NDVI. Finally, the effect of the prior was discussed.

In accordance with the philosophy of a Bayesian approach, efforts should be made towards: (1) improving the forward model that relates Tb to sm and τ and, (2) improving the *prior* pdf in order to constrain sm and τ retrievals. The latter can be enhanced by developing a deeply analysis on Aquarius RVI in order to fully exploit the combination of Aquarius active and passive L-band instruments, specially facing the upcoming of SMAP mission [**?**], and by introducing a more informative prior on sm.

Finally, it is important to consider the main concern of this approach: time performance. Although this limitation can be overcome through sampling methodologies such as MCMC method, the runtime is still large to drive operational global sm and τ retrievals. Nevertheless, further advances on computation methodologies will allow to derive an almost real time retrieval for low resolution systems. In addition, it constitutes a robust method for evaluation purposes.