Research Article

Research on Fusing Multisatellite Soil Moisture Data Based on Bayesian Model Averaging

Shan Wang ¹, Yuexing Wang,¹ Chi Zhang,¹ Han Shuai,² and Chun-Xiang Shi²

¹School of Information Engineering, East China Jiaotong University, Nanchang 330013, China

²National Meteorological Information Center, Beijing 100081, China

Correspondence should be addressed to Shan Wang; patrick_shan@163.com

Received 12 February 2018; Revised 13 May 2018; Accepted 15 May 2018; Published 25 June 2018

Academic Editor: Jifu Yin

Copyright © 2018 Shan Wang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Soil moisture (SM) is an important physical quantity that can reflect the land surface condition. There are many ways to measure SM, satellite microwave remote sensing is now considered the primary method because it can provide real-time high-resolution data. However, SM data obtained by satellite remote sensing exhibit certain deviation compared with reference data obtained from ground stations. To improve the accuracy of SM forecasts, this study proposed the use of a Bayesian model averaging (BMA) method to integrate multisatellite SM data. First, China was divided into eight regions. Then, SM data observed by satellites (FY3B, SMOS, and WINDSAT) were fused using the BMA method and a traditional averaging method. Finally, SM data were predicted using data from ground observation stations as a reference standard. Following the fusion process, three parameters (standard deviation, correlation coefficient, and root mean square deviation) were used to evaluate the fusion results, which revealed the superiority of the BMA method over the traditional averaging method.

1. Introduction

Soil moisture (SM) is an important parameter in surface climate and numerical weather forecasting, hydrological forecasting, agricultural drought monitoring, and the predictions of land surface models in the field of research. Therefore, obtaining high-quality SM data is important with regard to these various activities [1–3]. There are many ways to obtain SM data, such as ground-based observation and satellite microwave remote sensing [4]. Currently, because the distribution of ground observation stations is sparse, satellite microwave remote sensing has become the primary method to obtain real-time high-resolution SM data. Microwave sensors mounted on satellites can obtain large quantities of information on shallow SM. This information has been used in research regarding the validation of the inversion of satellite-derived SM products, as well as in applications of land surface models and numerical forecasting models [5, 6].

Since 1978, several satellite-borne active and passive microwave sensors have been launched successfully by China and other countries. For example, microwave sensors for detecting SM have been deployed on Aqua/AMSR-E,

Coriolis/WINDSAT, MetOp-A/ASCAT, SMOS/MIRAS, and FY3B/MWRI satellites. However, because of the influence of land surface roughness, vegetation, and the inversion algorithm adopted, SM data derived from satellite remote sensing often show certain deviation [7]. Therefore, how to integrate multisource remote sensing satellite data to achieve better results is a topic worth pondering. The traditional method of fusing SM data takes the average values of the observational data of each satellite for the fusion results. In other words, each satellite model is assigned the same weighting. However, in practice, each satellite model behaves differently in different regions [8]. Taking China as an example, SM derived by FY3B shows two extreme states: moisture in the humid areas of northern and southern China and dryness in arid northwestern areas. SM derived by SMOS is mostly dry in most parts of the country, and differences in its spatial distribution are not obvious. The soil moisture in WINDSAT is relatively dry in the whole country, but it is relatively moist in the middle and lower reaches of the Yangtze River and in the Northeast [9]. In addition, the traditional fusion method does not incorporate observational data from ground-based stations.

The objective of this study was to improve the accuracy of SM forecasts using a Bayesian model averaging (BMA) method to obtain time-varying weights. The a posteriori probability weights assigned to each sensor model in a specific period reflect the inherent uncertainty of each sensor model. In the early stages of the development of the BMA method, there have been some related research applications [10, 11]. The first application of the BMA method was to calibrate forecast ensembles, where the sea level pressure and other weather variables obeyed Gaussian distributions [12]. In 2007, the BMA method was applied to the variable of precipitation, which did not obey a Gaussian distribution [13]. In 2013, the BMA method was applied to the prediction of daily mean temperature in the Huaihe River Basin, China. A BMA probabilistic forecasting model for each site in the basin was established dynamically using the regional forecasting techniques of the TIGGE multi-model super ensemble forecasting system [14].

The average daily soil volumetric water content of 376 sites of the automatic soil moisture observation station (ASM) in 2012 was used in this study as a reference standard. This paper mainly studies the surface soil moisture data (0–10 cm). Three passive microwave remote sensing satellites (FY3B, SMOS, and WINDSAT) are used for satellite data [15, 16]. SM data observed by satellites were fused by the BMA method. The remainder of this paper is organized as follows. Section 2 introduces both the theory of the BMA method and the algorithm adopted. In Section 3, the method for evaluating the BMA model is described. Section 4 presents the original data and the SM forecast values adopting each region as the research object. Finally, our conclusions are stated in Section 5.

2. Principles of the BMA Method

2.1. Bayesian Model Averaging Method. Bayesian theory provides a set of ideas based on probabilistic statistical methods applicable to the fusion of information from different sources. The Bayesian formula is defined as

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}. \quad (1)$$

The BMA method is a statistical postprocessing method based on the Bayesian theory that uses a combination of multiple statistical models to produce a prediction. If y is the forecast variable (the SM data after fusion), $f = f_1, \dots, f_K$ is the prediction result of K possible models (the prediction result of K satellites), and y^T is the training data (the pre-real satellite remote sensing SM observational data), the BMA prediction model can be written as follows:

$$p\left[y|(f_1, \dots, f_K, y^T)\right] = \sum_{K=1}^K p(f_K|y^T) p_K\left(y|(f_K, y^T)\right), \quad (2)$$

where $p(f_K|y^T)$ is the weight value of BMA, which represents the prediction value of the K th satellite model as the a posteriori probability of the optimal prediction result. Greater weights mean higher accuracy of the prediction

result, and the sum of all model weights is 1. The term $p_K(y|f_K, y^T)$ represents the probability density function (PDF) of the predicted variable y for a given sample and model condition. For each possible model, it is necessary only to consider the proportion that it occupies throughout the prediction process.

The synthetic prediction of variable y based on the BMA method is based on using the probability $p(f_K|y^T)$ as the weight, and the PDF $p_K(y|f_K, y^T)$ of all the models is weighted to realize the probability prediction of the variables. In the study of satellite remote sensing of SM data, $p_K(y|f_K, y^T)$ can be regarded as a normal distribution function, and its prediction is expected to be a simple linear function $u_K + v_K f_K$ of the single prediction result. The variance is σ^2 , and $y|(f_K, y^T) \sim N(u_K + v_K f_K, \sigma^2)$ where u_K and v_K can be calculated by the linear regression method. Therefore, we can obtain the expectation of the BMA forecast as

$$E\left[y|(f_1, \dots, f_K, y^T)\right] = \sum_{K=1}^K w_K (u_K + v_K f_K). \quad (3)$$

At this point, the predicted value is a definite value, which can be compared with the predicted value of the single model. We denote space and time by subscripts s and t , such that f_{Kst} denotes the K th forecast in the ensemble for location s and time t [12, 17]:

$$\begin{aligned} \text{var}\left[y_{st}|\left(y_{1st}, \dots, y_{Kst}, y^T\right)\right] \\ = \sum_{K=1}^K w_K \left((u_K + v_K f_{Kst}) - \sum_{j=1}^K (u_j + v_j f_{jst}) \right)^2 + \sigma^2. \end{aligned} \quad (4)$$

The solution of parameter w_K and σ^2 in the upper model is the key. First, the maximum likelihood estimation method is used to determine the maximum value and then the expectation-maximization (EM) algorithm is used to solve the problem (Section 2.2 for details of the specific algorithms).

2.2. Parameter Estimation. The EM algorithm is a method for obtaining maximum likelihood estimations of parameters [18, 19]. Raftery et al. [12] proposed a new method to solve the weights and the variance using the EM algorithm to solve the case where the forecast variable obeys a normal distribution. The EM algorithm is iterative, and it alternates between two steps: the E (or expectation) step and the M (or maximization) step [20]. The two steps of the EM algorithm are as follows:

First, the weights and variances are initialized:

$$\begin{aligned} j &= 0, \\ w_K^{(j)} &= \frac{1}{K}, \end{aligned} \quad (5)$$

$$\sigma^{2(j)} = \frac{1}{n} \sum_{s,t} \sum_{K=1}^K [y_{st} - (u_K + v_K f_{Kst})]^2.$$

Its logarithmic likelihood function is

$$L(w_1^{(0)}, \dots, w_K^{(0)}, \sigma^{2(0)}) = \sum_{s,t} \log \left(\sum_{k=1}^K w_k^{(0)} g_K(u_K + v_K f_{Kst}, \sigma^{2(0)}) \right). \quad (6)$$

E step: for each j , replace it with $j+1$ and calculate

$$Z_{Kst}^{(j)} = \frac{w_K g_K(u_K + v_K f_{Kst}, \sigma^{2(j-1)})}{\sum_{l=1}^K w_l g_l(u_l + v_l f_{lst}, \sigma^{2(j-1)})}. \quad (7)$$

M step: its weight value is

$$w_K^{(j)} = \frac{1}{n} \sum_{s,t} Z_{Kst}^{(j)}. \quad (8)$$

We can then obtain the variance

$$\sigma^{2(j)} = \frac{1}{n} \sum_{s,t} \sum_{K=1}^K Z_{Kst}^{(j)} [y_{st} - (u_K + v_K f_{kst})]^2. \quad (9)$$

The above steps are repeated to update the iteration, constantly optimizing the parameter value and checking for convergence. The iteration continues until convergence is achieved.

3. Evaluation Criteria

The SM data obtained from the automatic ground observation stations were used as the reference standard, and the multisatellite SM data were recorded as x_i , the mean value of which was $\bar{x}_i$. The calibrated SM data were recorded as y_i , the mean value of which was $\bar{y}_i$. To evaluate the SM data, we introduce three parameters to measure the calibration results, which are the standard deviation (SD), correlation coefficient (R), and root mean square deviation (RMSD), using the SM data from the ground observation stations as the reference standard. For ground-observed SM data and satellite-derived SM data, the SD has the following formula:

$$SD = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2}. \quad (10)$$

The value of R between the corrected SM data and the reference SM data can be calculated as follows:

$$R = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(1/N) \sum_{i=1}^N (x_i - \bar{x})^2} \sqrt{(1/N) \sum_{i=1}^N (y_i - \bar{y})^2}}, \quad (11)$$

and the RMSD can be calculated as

$$RMSD = \sqrt{\frac{\sum_{i=1}^N [(x_i - \bar{x}) - (y_i - \bar{y})]^2}{N}}. \quad (12)$$

If either the SD or the RMSD is small or the R is closer to 1, the better result we will get. A Taylor diagram can characterize these three parameter values and illustrate the results more clearly (Section 4).

4. Regional SM Data Fusion Results

Unlike some studies, the SM data of each province were taken as the research object. However, it is more appropriate that the overall area was divided into several regions based on characteristics of drought and flooding. Zhu [21] used the rotary empirical orthogonal function to divide eastern and western regions of China into seven drought and flood areas. In this study, China was divided into eight regions, each of which was covered equally by three satellites. In the case of Northeast China, for example, we randomly selected 9/10 of all the data for training purposes. According to the Bayesian model, the EM algorithm was used to derive the weights of each satellite, and then the remaining 1/10 data were treated using the BMA method and the traditional averaging method to obtain the weights of the fusion and to draw the Taylor diagrams for comparison. Figures 1 and 2 depict the experimental results of the training for the Northeast region. Figure 1 describes the distribution of raw data in.

Figure 1 describes the distribution of raw data in Northeast China, including ground observation data and three satellite observations. The horizontal axis represents the distribution of the site, and the vertical axis indicates the value of soil moisture data. Through the curve trend of each data, it can be seen that the soil moisture has a strong spatial variability. It can be seen from the map that the trend of the WINDSAT satellite is the closest to the ground observation data (ASM), and the FY3B and SMOS satellite curves are far from the ASM value. Figure 2 illustrates the satellite observation data and the product of the BMA method after integration of the PDF curve, where the abscissa represents the soil moisture and the ordinate represents the soil volume of water content of the data. From the PDF curve describing the cumulative distribution, it is evident that WINDSAT is the closest of the three satellites to the fusion results and WINDSAT has the highest weight. In addition, we obtained three satellites with weights of 0.268 (FY3B), 0.211 (SMOS), and 0.521 (WINDSAT), respectively, in Northeast region.

To evaluate the quality of the results obtained by the fusion method, we used the parameters of SD, R , and RMSD to measure the system deviation and plotted them on a Taylor diagram for an intuitive representation. In the Taylor diagrams, the distance from the origin (the radius) represents the SD of the data. The arc represents the value of R between the point and the reference data. Drawing an arc with the reference point as its center, each point located on the arc represents the RMSD of that point. In the process of drawing the Taylor diagrams, some difficulties were encountered; however, the results were finally optimized using quality control.

Figures 3–5 present Taylor diagrams for Northeast China, northern North China, and the region of the middle and lower reaches of the Yangtze River, respectively. In figures, the six points plotted in the Taylor diagrams are respectively the automatic ground-based observational data (ASM), observational data from the three satellites (FY3B, SMOS, and WINDSAT), fusion results obtained by means of the traditional averaging method (AVER), and BMA results. As can be seen from Figure 3, the SD of the WINDSAT

FIGURE 1: Northeast region ground-based observational data (ASM) and remote sensing data from the three satellites (FY3B, SMOS, and WINDSAT).

FIGURE 2: Northeast region satellite data and product of the BMA method after integration of the PDF curve.

satellite in Northeast China is about 0.07, R is more than 0.6, and the RMSD is about 0.06. These three evaluation indexes are all superior to the SMOS satellite and the FY3B satellite, which corresponds to the information expressed in the previous results. Using the multisatellite data to fuse, the SD of the fusion results of the average method is about 0.075, R reaches 0.6, and the RMSD is about 0.06. And the SD of the fusion result through BMA method is 0.06, R reaches 0.7, and RMSD is 0.05. The parameters of the BMA method have been improved accordingly. The experimental results show that the BMA method can improve the accuracy of the prediction compared with the mean method and the BMA method has a better fusion effect.

In Figure 4, three parameters (SD, R , and RMSD) of the WINDSAT satellite in northern North China are superior to that of the SMOS satellite and the FY3B satellite. Using the multisatellite data to fuse, the standard deviation of the fusion results of the average method is larger than the standard deviation of the BMA method. The parameter of R

is approximately equal to that of the BMA method, and RMSD is greater than the root mean square error of the BMA method. It is still the better fusion effect of the BMA method. From Figure 5, we can see that the three evaluation indexes of satellite data are similar to those of the above two regions, while the results of the traditional mean method and the BMA method are similar. But we can see that the BMA method is still better. However, environmental and other factors led to insufficient data for the Northwest and Southwest regions, which produced unsatisfactory results. These areas have high altitude and poor weather, and these regional meteorological stations are scarce. This will be investigated in future research.

In order to compare the overall performance of the BMA method and the traditional average method in different regions, we give the results of multiple regions in a Taylor diagram. Figure 6 shows the results of the Taylor diagram of five regions. The regions 1 to 5 are, respectively, Northeast China, northern North China, southern North China, the

FIGURE 3: Taylor diagram for Northeast China.

FIGURE 5: Taylor diagram for the region of the middle and lower reaches of the Yangtze River.

FIGURE 4: Taylor diagram for northern North China.

FIGURE 6: Taylor diagram for five regions.

middle and lower reaches of the Yangtze river, and eastern Northwest China. In the figure, 1FY represents the result of FY3B in region 1; correspondingly, 1SM stands for the result of SMOS in region 1, and 1WI is the result of WINDSAT in region 1. From the figure, we can see that the points named BMA is located below the AVER points. It means that the SD and RMSD of the fusion results of the averaging method are greater than the BMA method. Furthermore, the R of the fusion results of BMA method is closer to 1. We can know that the BMA method is better than the traditional averaging method in the five regions.

Table 1 shows a comparison of the fusion characteristics based on region division. The first three columns of the table are the previous divided area and the divided area of this paper, and the corresponding latitude and longitude of each area. The next two columns are quantity of data fusion, and BMA method fusion effect is better than the averaging method number of sites. The last column is ratio of better

result in BMA method. It can be seen that the proportion of BMA method fusion in the Northeast, eastern Northwest, and southern Southwest regions is 78.96%, 57.33%, and 92.46%, respectively, whereas it is <50% in all other regions. By querying the satellite weight in those regions with values <50%, it was found that the weight value of the BMA method is about 1/3, similar to the weight value obtained by the averaging method. The small amounts of data obtained for these regions because of environmental and other factors account for the poor results. In spite of this, the overall result of the fusion of BMA methods is clearly superior to the traditional average method.

5. Conclusions

This research used the BMA method for the fusion of multisatellite microwave remote sensing SM data, in order to

TABLE 1: Comparison of regional divisions and fusion characteristics.

Division by Zhu	Division of this paper	Latitude and longitude	Quantity of data fusion	Better result in BMA	Ratio of better result in BMA (%)
Northeast China	Northeast China	120°E-135°E 40°N-50°N	556	439	78.96
Northern North China	Northern North China	110°E-120°E 40°N-45°N	637	269	42.23
Southern North China	Southern North China	110°E-120°E 34°N-40°N	2100	974	46.38
Northwest China	Eastern Northwest China	95°E-110°E 34°N-42°N	1153	661	57.33
	Western Northwest China	80°E-95°E 34°N-50°N	—	—	—
	The middle and lower reaches of the Yangtze river	110°E-122°E 30°N-34°N	868	353	40.67
Southwest China	Northern Southwest China	100°E-110°E 28°N-34°N	445	202	45.39
	Southern Southwest China	100°E-110°E 20°N-28°N	623	576	92.46

improve the prediction of SM data at unmeasured points. From the results of Taylor diagram, the fusion results obtained by using BMA method in different regions are better than the traditional average method. From the fusion results of different regions, it can be known that the proportion of the BMA method fusion in the Northeast region is better, accounting for 78.96%, the eastern Northwest region reaching 57.33%, and the southern Southwest region reaching 92.46%. So we can get the following conclusions. In the prediction of SM data, the use of the BMA method for the fusion of SM data not only solves the uncertainty of the model but it also improves the accuracy of the predicted value.

Although the evaluation parameters were improved, it has not been clarified how the forecast could be made more accurate. In future study, we will use additional ground-based SM observation stations and more sophisticated equipment to obtain further SM data to augment the original database.

Data Availability

The data used to support the findings of this study are included within the supplementary information (in txt format).

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this article.

Acknowledgments

The authors not only wish to acknowledge the National Meteorological Center for providing the data, but also thank teacher Shi for her guidance and support. This work was supported in part by Major Program of National Natural Science Foundation of China (no. 91437220), Jiangxi Province Science Foundation for Youths (no. 20171ACB21038), and JiangXi Municipal Science and Technology Project.

Supplementary Materials

The supplementary materials include the data observed from three satellites (FY3B, SMOS, and WindSat) and 376 automatic soil moisture stations (ASM, 0–10 cm) daily data in 2012, and all data are arranged according to the regions. (*Supplementary Materials*)

References

- [1] C. Haishan and Z. Jing, "Impact of interannual soil moisture anomaly on simulation of extreme climate events in China. Part II: sensitivity experiment analysis," *Chinese Journal of Atmospheric Sciences*, vol. 37, no. 1, pp. 1–13, 2013, (in Chinese).
- [2] M. Zhuguo, F. Congbin, X. Li et al., "Some problems in the study on the relationship between soil moisture and climatic change," *Advance in Earth Sciences*, vol. 16, no. 4, pp. 563–568, 2001, (in Chinese).
- [3] S. Shuanghe, J. Long, Z. Ying et al., "A practical simulation model for cropped soil water prediction," *Scientia Meteorologica Sinica*, vol. 16, no. 3, pp. 240–248, 1996, (in Chinese).
- [4] R. D. Jeu and W. Dorigo, "On the importance of satellite observed soil moisture," *International Journal of Applied Earth Observation and Geoinformation*, vol. 45, pp. 107–109, 2016.
- [5] W. Wagner, G. Lemoine, and H. Rott, "A method for estimating soilmoisture from ERS scatterometer and soil data," *Remote Sensing of Environment*, vol. 70, no. 2, pp. 191–207, 1997.
- [6] E. G. Njoku, T. J. Jackson, V. Lakshmi, T. K. Chan, and S. V. Nghiem, "Soil moisture retrieval from AMSR-E," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 41, no. 2, pp. 215–229, 2003.
- [7] M. Parrens, J. P. Wigneron, P. Richaume et al., "Considering combined or separated roughness and vegetation effects in soil moisture retrievals," *International Journal of Applied Earth Observation and Geoinformation*, vol. 55, pp. 73–86, 2017.
- [8] T. Sakai, T. Iizumi, M. Okada et al., "Varying applicability of four different satellite-derived soil moisture products to global gridded crop model evaluation," *International Journal of*

- Applied Earth Observation and Geoinformation*, vol. 48, pp. 51–60, 2016.
- [9] Y. Zhuang, C. Shi, R. Shen et al., “Quality evaluation of multi-microwave remote sensing soil moisture products over China,” *Journal of the Meteorological Sciences*, vol. 35, no. 3, pp. 289–296, 2015.
 - [10] R. E. Kass and A. E. Raftery, “Bayes factors,” *Journal of the American Statistical Association*, vol. 90, no. 430, pp. 773–795, 1995.
 - [11] J. A. Hoeting, D. Madigan, A. E. Raftery et al., “Bayesian Model Averaging: a Tutorial,” *Statistical Science*, vol. 14, no. 4, pp. 382–401, 1999.
 - [12] A. E. Raftery, T. Gneiting, F. Balabdaoui, and M. Polakowski, “Using Bayesian model averaging to calibrate forecast ensembles,” *Monthly Weather Review*, vol. 133, no. 5, pp. 1155–1174, 2005.
 - [13] J. M. Sloughter, A. E. Raftery, and T. Gneiting, “Probabilistic quantitative precipitation forecasting using Bayesian model averaging,” Tech. Rep. 496, Department of Statistics, University of Washington, Seattle, WA, USA, 2006.
 - [14] J. Liu, Z. Xie, and L. Zhao, “BMA probabilistic forecasting for the 24-h TIGGE multi-model ensemble forecasts of surface air temperature,” *Chinese Journal of Atmospheric Sciences*, vol. 37, no. 1, pp. 43–53, 2013.
 - [15] Z. F. Zhang, “Development of quality control program for soil moisture data of automatic station,” *Journal of Arid Land Geography*, vol. 36, no. 1, pp. 101–108, 2013.
 - [16] Z. Dongbin, “Automatic soil moisture observation data quality analysis and reference site selection,” Technical Report, National Meteorological Information Center, Beijing, China, 2013, (in Chinese).
 - [17] A. E. Raftery, “Bayesian model selection in structural equation models,” in *Testing Structural Equation Models*, K. A. Bollen and J. S. Long, Eds., SAGE Publications, Thousand Oaks, CA, USA, 1993.
 - [18] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of the Royal Statistical Society Series B*, vol. 39, no. 1, pp. 1–38, 1977.
 - [19] G. J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*, Wiley, Hoboken, NJ, USA, 1997.
 - [20] C. F. J. Wu, “On the convergence properties of the EM algorithm,” *Annals of Statistics*, vol. 11, no. 1, pp. 95–103, 1983.
 - [21] Z. Yafen, “Zhu Yafen’s drought and flood division in eastern China and evolution of drought and flood in North China during 5.30,” *Journal of Geography*, vol. 58, no. S1, pp. 100–107, 2003.