



This is a repository copy of *Soil moisture deficit estimation using satellite multi-angle brightness temperature*.

White Rose Research Online URL for this paper:
<http://eprints.whiterose.ac.uk/155713/>

Version: Accepted Version

Article:

Zhuo, L. orcid.org/0000-0002-5719-5342, Han, D. and Dai, Q. (2016) Soil moisture deficit estimation using satellite multi-angle brightness temperature. *Journal of Hydrology*, 539. pp. 392-405. ISSN 0022-1694

<https://doi.org/10.1016/j.jhydrol.2016.05.052>

Article available under the terms of the CC-BY-NC-ND licence
(<https://creativecommons.org/licenses/by-nc-nd/4.0/>).

Reuse

This article is distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs (CC BY-NC-ND) licence. This licence only allows you to download this work and share it with others as long as you credit the authors, but you can't change the article in any way or use it commercially. More information and the full terms of the licence here: <https://creativecommons.org/licenses/>

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.



eprints@whiterose.ac.uk
<https://eprints.whiterose.ac.uk/>

Soil moisture deficit estimation using satellite multi-angle brightness temperature

Lu Zhuo^{1*}, Dawei Han¹, Qiang Dai²

¹WEMRC, Department of Civil Engineering, University of Bristol, Bristol, UK

²Key Laboratory of Virtual Geographic Environment of Ministry of Education, School of Geography Science,

Nanjing Normal University, Nanjing, China

*Correspondence: lu.zhuo@bristol.ac.uk

Abstract

Accurate soil moisture information is critically important for hydrological modelling. Although remote sensing soil moisture measurement has become an important data source, it cannot be used directly in hydrological modelling. A novel study based on nonlinear techniques (a local linear regression (LLR) and two feedforward artificial neural networks (ANNs)) is carried out to estimate soil moisture deficit (SMD), using the Soil Moisture and Ocean Salinity (SMOS) multi-angle brightness temperatures (T_{bs}) with both horizontal (H) and vertical (V) polarisations. The gamma test is used for the first time to determine the optimum number of T_{bs} required to construct a reliable smooth model for SMD estimation, and the relationship between model input and output is achieved through error variance estimation. The simulated SMD time series in the study area is from the Xinanjiang hydrological model. The results have shown that LLR model is better at capturing the interrelations between SMD and T_{bs} than ANNs, with outstanding statistical performances obtained during both training (NSE = 0.88, $r = 0.94$, RMSE = 0.008 m) and testing phases (NSE = 0.85, $r = 0.93$, RMSE = 0.009 m). Nevertheless, both ANN training algorithms (radial BFGS and conjugate gradient) have performed well in estimating the SMD data and showed

24 excellent performances compared with those derived directly from the SMOS soil moisture
25 products. This study has also demonstrated the informative capability of the gamma test in the
26 input data selection for model development. These results provide interesting perspectives for data-
27 assimilation in flood-forecasting.

28 **Keywords:** SMOS brightness temperature; soil moisture; local linear regression (LLR); artificial
29 neural networks (ANNs); soil moisture deficit (SMD); hydrological modelling

30 **1. Introduction**

31 Although soil moisture comprises only 0.01% of the total amount of water on our planet, its
32 existence plays an important role in influencing the water and energy exchanges at the land
33 surface/atmosphere interface. There is abundant evidence that hydrological processes are
34 significantly conditioned by a river catchment's antecedent wetness state ([Massari et al., 2014](#);
35 [Tramblay et al., 2012](#)). In particular the surface soil wetness is an important variable in
36 hydrological modelling because it controls key processes such as runoff and evapotranspiration,
37 and is a vital parameter for flood modelling ([Draper et al., 2011](#); [Han et al., 2012](#))

38 The Earth thermal emission at microwave bands depends essentially on the soil temperature and
39 the soil water content ([Al-Yaari et al., 2014](#); [Rodríguez-Fernández et al., 2015](#)). Recent research
40 activities indicate rising interest in the operational monitoring of the global soil moisture remote
41 sensing. In particular, the data acquired by lower microwave frequencies (e.g., L-band at 1.20-1.41
42 GHz), both active and passive, have been utilised to provide detailed surface soil moisture
43 fluctuations in recent years ([Calvet et al., 2011](#)). The launch of the Soil Moisture and Ocean
44 Salinity (SMOS; ([Kerr et al., 2001](#))) mission in November 2009 and the Soil Moisture
45 Active/Passive mission (SMAP; ([Entekhabi et al., 2010](#))) in January 2015 clearly demonstrates the

46 significance and determination of an advanced global surface soil moisture monitoring system.
47 SMOS is the first mission dedicated to monitoring direct surface soil moisture and sea surface
48 salinity on a global scale ([Kerr et al., 2010](#)), and has a longer period of data record since its launch
49 in 2009. Therefore, SMOS is chosen in this study.

50 The SMOS soil moisture operational algorithm utilises a direct or forward model and an optimal
51 estimation method: a radiative transfer model (e.g., LMEB model is used in the SMOS algorithm
52 ([Wigneron et al., 2007](#))) is applied to estimate L-band brightness temperatures (hereafter T_{bs}) for
53 a set of physical parameters, soil composition, and moisture content and vegetation opacity
54 ([Rodríguez-Fernández et al., 2015](#)). In order to estimate soil moisture, the simulated T_{bs} are
55 compared with those measured by SMOS using an iterative process to minimise the difference
56 between them. This approach then requires in-situ observation data for soil moisture evaluation
57 ([Al-Yaari et al., 2014](#); [Al Bitar et al., 2012](#)). However most areas do not have in-situ sensors
58 because they are expensive to set up and impractical to maintain; and they are too sparse for
59 catchment-scale studies ([Al - Shrafany et al., 2013](#); [Srivastava et al., 2013b](#); [Srivastava et al.,](#)
60 [2013c](#); [Walker et al., 2004](#); [Wang and Qu, 2009](#)). Since the presence of vegetation can reduce the
61 brightness temperature sensitivity to soil moisture, in the aforementioned method decoupling the
62 effects of soil and vegetation on brightness temperature can pose a major challenge for useful
63 application under such circumstances.

64 In order to retrieve accurate soil wetness information that can be directly used in a hydrological
65 model and avoid aforementioned shortcomings, a data-driven model is desirable, which could
66 effectively link the inputs to the desired output and is not computationally intensive. This can be
67 achieved by building an inverse model that provides soil moisture information (i.e., soil moisture
68 deficit (SMD) in this study, which is a key soil moisture variable in hydrological models ([Zhuo et](#)

69 [al., 2015a](#)) directly from a given set of satellite measured T_{bs} . Among the data-driven models,
70 nonlinear regression models such as Local Linear Regression (LLR) and Artificial Neural
71 Networks (ANNs) are widely recognised and used as efficient inverse models. Therefore both LLR
72 and ANNs are used in this study.

73 The foremost objective of this study is therefore to build an inverse model for the first time that
74 can simulate the relevant hydrological SMD data directly from the SMOS brightness temperatures
75 using various nonlinear modelling techniques. In this study, the SMD is estimated instead of the
76 normal soil moisture because in hydrological modelling the excess runoff is closely linked with
77 SMD, but not directly with the normal soil moisture (i.e., the volumetric soil moisture). The SMD
78 refers to the amount of water needed to bring the soil moisture back to field capacity. Since SMD
79 is directly relevant to hydrology, it is the main purpose of this study. SMOS is the first radiometer
80 in space with full-polarisation and multangular capabilities ([Rodríguez-Fernández et al., 2015](#)).
81 Hence, a dedicated retrieval scheme has to be studied. An LLR model and two ANN models are
82 trained and tested for their valuation in SMD retrieval. The modelled SMD values using different
83 techniques are then compared against the Xinanjiang simulated SMD as the target. Furthermore, a
84 well-proven and widely applied computing algorithm called the gamma test (GT) is employed to
85 find the optimal combination of data inputs for SMD calculation. [Noori et al. \(2011\)](#) and [Remesan](#)
86 [et al. \(2008\)](#) applied the GT data selection method in hydrological studies, for daily solar radiation
87 estimation and monthly streamflow prediction, and both reported positive performances. In
88 contrast to the conventional allocation method of the training and the testing data, the M-test is
89 adopted to find the optimal training dataset which has sufficient information for training any
90 regression models. This will avoid wasting time and effort in allocating excessive training data or
91 using inadequate training data. Therefore, no predefined training and testing data will be specified

92 at the early stage of the study. Finally, the SMD estimates from the aforementioned nonlinear
93 methods are compared with those directly derived from the SMOS soil moisture products (i.e., two
94 different SMOS products are used: one is from the SMOS Barcelona Expert Centre (SMOS-BEC)
95 ([SMOS-BEC, 2015](#)) and the other is from the Centre Aval de Traitement des Données SMOS
96 (SMOS-CATDS) ([Jacquette et al., 2010](#))).

97 **2. Study area and data**

98 Pontiac is a medium-sized catchment (1500 km²) in the Vermilion River, located in the central
99 Illinois area of the U.S. The catchment's topography is flat and mainly used for cultivation purpose
100 as illustrated in Fig. 1b ([Bartholomé and Belward, 2005](#); [Hansen, 1998](#)). Based on the Global Soil
101 Regions map ([USDA, 2005](#)), its soil is predominately Mollisols. The catchment is dominated
102 mainly by hot summer continental climate ([Peel et al., 2007](#)). The layout of the Pontiac catchment
103 is shown in Fig. 1a along with the location of its flow gauge, river network, and the North American
104 Land Data Assimilation Systems Phase 2 (NLDAS-2) grid points (i.e., the marked grid points are
105 located at the central of each 0.125° x 0.125 ° NLDAS-2 grids). The spatial variations of an
106 extracted SMOS T_b dataset (H polarisation) at an incidence angle of 32.5° is shown in Fig. 1c (it
107 has been transformed into NLDAS-2 grid spacing at 0.125° for easier analysis). It can be seen from
108 this retrieved image, the central catchment area has lower T_b values (i.e., relatively wetter soil),
109 while the western upper and lower parts show slightly higher T_b values (i.e., relatively drier soil).
110 This could partially be explained by the location of the river network as indicated in Fig. 1a: the
111 majority of the water concentrates at the central area (i.e., the mainstream) and then flows to the
112 catchment outlet (so the soil can be replenished with water more easily); whereas the soil around
113 the small substream areas has less water availability and tends to be drier. It should be noted that

114 soil moisture does not solely correlate with the variation of brightness temperature but also with
115 other factors such as vegetation cover, local soil properties, and surface roughness.

116 The Xinanjiang (XAJ) model's hydrological forcing is obtained from the NLDAS-2 ([Mitchell et](#)
117 [al., 2004](#)). The datasets comprise precipitation ([Daly et al., 1994](#)) and potential evapotranspiration
118 at the 0.125° spatial resolution and daily temporal resolution (converted from hourly resolution).
119 Both datasets have been transformed into the catchment-scale using the weighted average method
120 to operate the lumped XAJ model. Readers are referred to [Xia et al. \(2012\)](#) and [Zhuo et al. \(2015c\)](#)
121 for a full description of the NLDAS-2 data products. The observed daily flow data for this study
122 is provided by the U.S. Geological Survey. The observations cover a total period of 24-months
123 from January 2010 to December 2011. The reason for using these two-year data is due to the
124 discontinuity of flow observations in the selected catchment.

125 **2.1 SMOS data**

126 SMOS retrieves the thermal emission from the Earth at the frequency of 1.4 GHz in both
127 polarisations and for incidence angles from 0° to 60°. It is dedicated to providing global surface
128 soil moisture information at an accuracy of 0.04 m³/m³ ([Kerr et al., 2012](#)). SMOS has a Y-shaped
129 antenna structure, which comprises 69 small antennas (a diameter of 16.5 cm) and 4.5-m long arms
130 to perform interferometry and synthesise an aperture of ~ 7.5 m ([McMullan et al., 2008](#);
131 [Rodríguez-Fernández et al., 2015](#)). The projection of the synthesised beam on the Earth surface is
132 generally presented as an ellipse whose axis ratio and orientation depend on the observed point
133 position ([Rodríguez-Fernández et al., 2015](#)). The retrieved observations have a spatial resolution
134 of 35-50 km ([Kerr et al., 2010](#)). SMOS follows a sun-synchronous polar orbit with a global

135 coverage at the equator crossing the times of 6:00 A.M. at the local solar time (LST) (ascending)
136 and 6:00 P.M. (LST, descending).

137 In order to estimate SMD from SMOS T_{bs} , the Level-3 brightness temperature data from the
138 CATDS is used ([Jacquette et al., 2010](#)). This daily global brightness temperature data contains
139 SMOS T_{bs} in the reference frame of 0.25° EASE grid ([Brodzik and Knowles, 2002](#)) on the Earth
140 surface. It provides T_{bs} measurements acquired at all incidence angles in a given day (averaged in
141 5° -width angle bins) which have been transformed into the ground polarisation reference frame
142 (i.e., H, and V polarisations). Hence, the quantity of the input data can be as high as 24 (12 angle
143 bins per polarisation), with the centre of the first angle bin at 2.5° in both polarisations ([Rodriguez-
144 Fernandez et al., 2014](#)). In this catchment, the only angle range that gives the most available record
145 of data is from 27.5° to 57.5° (i.e., 7 for H and 7 for V polarisation), which is therefore chosen for
146 the model development. In order to better understand the sensitivity of SMOS T_{bs} to the SMD, the
147 Pearson correlation coefficients (r) are calculated and illustrated in Fig. 2. It can be seen that the
148 correlation decreases for H polarisation when the incidence angle rises (from $r = \sim 0.55$ to $r = \sim$
149 0.45); whereas the correlation for V polarisation is more stable and fluctuates around $0.6 - 0.65$.
150 This phenomenon agrees with the general trend of the theoretical effect of H-V polarisations at
151 different incidence angles ([Wei et al., 2014](#)).

152 Additionally, the Level-3 soil moisture products from the CATDS (SMOS-CATDS) and the BEC
153 (SMOS-BEC) are also obtained for a comparison study. The main difference between these two
154 products is that they are made from different data inputs. The SMOS-BEC utilises the Level-2 Soil
155 Moisture User Data Product (UDP) generated by ESA as its Level-3 data inputs, while SMOS-
156 CATDS goes in a rather unusual way by using brightness temperature products in the Fourier
157 domain (L1B) as input for the Level-3 processor. The detailed comparison between these two

158 products is beyond the scope of this paper, and the interested readers are referred to [Elsa et al.](#)
159 [\(2013\)](#) and [SMOS-BEC \(2015\)](#) for full descriptions. All acquired SMOS products cover the period
160 between January 2010 and December 2011 and have been converted into a catchment-scale dataset
161 by the weighted average method. Furthermore, they have been re-scaled by mapping the mean to
162 zero and the standard deviation to 0.5. This normalisation step is able to equalise the relative
163 numerical difference among the input variables and better aid the GT feature selection routine
164 [\(Remesan et al., 2008\)](#). It is noted that the re-scaled data is only for the GT routine and the M-test,
165 and normal data are used for SMD estimation.

166 **3. Methodology**

167 **3.1 XAJ model**

168 The XAJ model developed by [Zhao \(1980, 1992\)](#) and [Zhao and Liu \(1995\)](#) is a widely used
169 conceptual rainfall-runoff model. The model has been proven in many publications to be effective
170 for both operational and offline simulation purposes in humid, semi-humid regions ([Chen et al.,](#)
171 [2013](#); [Shi et al., 2011](#); [Zhao, 1992](#); [Zhao and Liu, 1995](#); [Zhuo et al., 2015b](#); [Zhuo et al., 2015c](#)) as
172 well as dry areas ([Gan et al., 1997](#)) around the world. The main hypothesis used in the model
173 development is the runoff generation on repletion of its storage capacity, which means that runoff
174 is not generated until the soil water reaches the field capacity ([Zhao, 1992](#)). In this study, the XAJ
175 model is used for SMD estimation through an improved soil moisture accounting scheme ([Zhuo](#)
176 [and Han, 2016a,b](#)). Further details on calibration and validation of the XAJ model and the SMD
177 are discussed by [Zhuo et al. \(2015a\)](#) and [Zhuo et al. \(2016\)](#).

178 **3.2 Gamma test and M-test**

179 An appropriate selection of the incidence angles of the SMOS observations is important to ensure
180 the best SMD estimation. In this study, a well-developed GT algorithm ([Koncar, 1997](#); [Stefánsson
181 et al., 1997](#)) is adopted because it has been proven to be efficient in selecting model inputs ([Durrant,
182 2001](#); [Jaafar and Han, 2011](#); [Noori et al., 2011](#); [Remesan et al., 2008](#); [Tsui et al., 2002](#)). It is a
183 near-neighbour data analysis routine which allows efficient estimation of the minimum mean-
184 squared error (MSE) that can be achieved when modelling the input-output data using nonlinear
185 models. This calculation is called the gamma statistics and represented as Γ . The inspiration of GT
186 came from the Delta test ([Pi and Peterson, 1994](#)). Only a brief introduction on GT is provided here
187 and the interested readers are referred to the aforementioned papers for further explanations. For
188 simplicity a case is introduced where a set of data samples is given in the form of:

$$189 \quad \{ (x_i, y_i), 1 \leq i \leq M \} \quad (1)$$

190 where the input vectors $x_i \in \mathbb{R}^m$ are confined to a closed bounded set $C \in \mathbb{R}^m$, and without loss of
191 generality, the outputs $y_i \in \mathbb{R}$ are scalars. The vectors x comprise predictively useful information
192 that controls the output y . The only assumption made is that the underlying relationship of the
193 system is from the following equation:

$$194 \quad y = f(x_1 \dots x_m) + r \quad (2)$$

195 where f is a smooth function and r is an indeterminable variable that is regarded as noise. Without
196 loss of generality, the mean of the r distribution is assumed to be zero (because any constant bias
197 has been considered in the unknown function f) and that the variance of the noise $\text{Var}(r)$ is bounded.
198 The domain of a potential model is now restricted to the class of smooth functions which have

199 bounded first partial derivatives. The Γ is an estimate of the model's output variance that cannot
 200 be accounted for by a smooth data model.

201 The GT is based on $N[i, k]$, which are the k th ($1 \leq k \leq p$) nearest neighbours $x_{N[i, k]}$ ($1 \leq k \leq p$) for
 202 each vector x_i ($1 \leq i \leq M$). p is a fixed integer. GT is calculated from the Delta function of the input
 203 vectors:

$$204 \quad \delta_M(k) = \frac{1}{M} \sum_{i=1}^M |x_{N[i, k]} - x_i|^2 \quad (1 \leq k \leq p) \quad (3)$$

205 where $|\dots|$ is Euclidean distance, and the related gamma function of the output values:

$$206 \quad \gamma_M(k) = \frac{1}{2M} \sum_{i=1}^M |y_{N[i, k]} - y_i|^2 \quad (1 \leq k \leq p) \quad (4)$$

207 where $y_{N[i, k]}$ is the corresponding output value with $x_{N[i, k]}$. To compute Γ a least-squared
 208 regression line for the p points $(\delta_M(k), \gamma_M(k))$ is built as in the following equation:

$$209 \quad \gamma = A\delta + \Gamma \quad (5)$$

210 where Γ is the intercept on the vertical axis (i.e., $\delta = 0$), as can be explained as:

$$211 \quad \gamma_M(k) \rightarrow \text{Var}(r) \text{ in probability as } \delta_M(k) \rightarrow 0 \quad (6)$$

212 This gives an estimation of the optimal MSE value achievable utilising a modelling method for
 213 unknown smooth functions. The derived gradient A is also a useful indicator in showing
 214 information on the complexity of the system under investigation (the larger the A value the more
 215 complexity the model is required). The merit of GT is that it can provide valuable guidance about
 216 the system regardless of the subsequent modelling technique choice. A formal mathematical proof

217 of the GT can be found in [Evans and Jones \(2002\)](#). In practice, the GT can be carried out through
218 the winGamma™ software ([Durrant, 2001](#)).

219 A general practice in nonlinear modelling (e.g., LLR and ANNs) is to divide the dataset into two
220 parts, i.e., training and testing. However many studies hastily adopted the size of their training
221 dataset without proper examination, and this could result in unsatisfactory modelling performance.
222 Therefore in order to determine the best training data size that can give a stable and reliable Γ
223 statistics, an M-test is carried out. The M-test is accomplished by computing the Γ for increasing
224 M value (indicating the effect of the training data size) and through analysing the resulting graph
225 to determine whether the Γ approaches a stable asymptote (this way is easier than defining a
226 complex algorithm). Such a procedure is useful in avoiding wasteful model-fitting attempts when
227 the MSE from the training phase is already smaller than the $\text{Var}(r)$, and hence preventing the
228 overfitting problem.

229 **3.3 Nonlinear models**

230 The modern statistical approach to nonlinear model building has led to techniques such as LLR,
231 support vector machines, principal component analysis, feedforward ANNs, and radial basis
232 function networks. In this study, the LLR and the ANNs are used. Only brief theoretical
233 backgrounds relevant to the study are explained.

234 **3.3.1 Local linear regression (LLR)**

235 LLR is a widely researched nonparametric regression methodology that has been applied in low-
236 dimensional forecasting and smoothing problems ([Liu et al., 2011](#); [Pinson et al., 2008](#); [Remesan
237 et al., 2008](#); [Sun et al., 2003](#)). However to our knowledge it has rarely been used in soil moisture
238 estimation, especially those simulated from the remote sensing technology. The advantages of

239 LLR are that it can locally provide reliable statistical modelling based on a small amount of data
 240 sample, is less computationally demanding, and is able to give accurate estimations in regions of
 241 high data density in the input space. Furthermore, LLR can make an initial prediction with only
 242 three data points, and any newly updated data are used for further predictions. LLR performs local
 243 linear regression through the p_{\max} nearest points to a query point, to give a linear model in the
 244 locality of the query point. This process is repeated across the training data to produce a piecewise
 245 linear model. One of the methods of choosing p_{\max} is called influence statistics and is explained
 246 below ([Durrant, 2001](#); [Remesan et al., 2008](#)).

247 Given a neighbourhood of p_{\max} points, the following linear matrix equation needs to be calculated

$$248 \quad X\mathbf{m} = \mathbf{y} \tag{7}$$

249 where X is a $p_{\max} \times d$ matrix of the p_{\max} input points in d dimensions, x_i ($1 \leq i \leq p_{\max}$) are the
 250 nearest neighbour points, \mathbf{y} is a column vector at the length p_{\max} of the associated outputs, and \mathbf{m} is
 251 a column vector of parameters that has to be determined to provide the best mapping solution from
 252 X to \mathbf{y} , such that

$$253 \quad \begin{pmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1d} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{p_{\max}1} & x_{p_{\max}2} & x_{p_{\max}3} & \cdots & x_{p_{\max}d} \end{pmatrix} \begin{pmatrix} m_1 \\ m_2 \\ \vdots \\ m_d \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{p_{\max}} \end{pmatrix} \tag{8}$$

254 The rank of the matrix X is the number of linearly independent rows, which affects the existence
 255 or uniqueness of the solutions for \mathbf{m} .

256 If the matrix X is square and non-singular then the unique solution to Equation (7) is $m = X^{-1}y$.
257 However if X is not square or singular, Equation (7) needs to be modified and m is determined by
258 minimising the following equation:

$$259 \quad |Xm - y|^2 \tag{9}$$

260 as has been proved by [Penrose \(1955\)](#), the distinct solution to this problem is:

$$261 \quad m = X^\# y \tag{10}$$

262 where $X^\#$ is a pseudo-inverse matrix ([Penrose, 1955](#); [Penrose, 1956](#)).

263 One of the various methods available to organise the input training data is the k-dimensional tree
264 (k-d tree), with a time complexity in the order $O(M \log M)$. A k-d tree is a space partitioning data
265 structure for organising points in a k-dimensional space so that the LLR algorithm can be
266 implemented using the least number of direct evaluations ([Remesan et al., 2008](#)).

267 **3.3.2 Artificial neural networks (ANNs)**

268 ANNs are models that learn from a training data set mimicking the human-learning ability ([Zurada,](#)
269 [1992](#)). They are able to identify noisy data and approximate multivariate nonlinear relations among
270 the variables ([Ahmad et al., 2010](#)). They have been widely used in many disciplines, including
271 water resources and hydrology research such as for river level forecasting, rainfall runoff
272 modelling, daily evaporation estimation, rainfall forecasting and groundwater modelling
273 ([Dehghani et al., 2014](#); [Han et al., 2007](#); [Ireland et al., 2015](#); [Islam et al., 2012](#); [Srivastava et al.,](#)
274 [2013a](#); [Tehrany et al., 2014](#)). Multilayer feedforward neural networks (NNs) are universal
275 approximators ([Hornik et al., 1989](#)) and explored in this study to determine their effectiveness in
276 relating a number of inputs to the SMD. Specifically, an ANN can exploit the synergy of different

277 input variables due to its truly multivariate nature and its nonlinear capabilities ([Aires et al., 2011](#)).

278 The supervised ANN is the most widely applied ANN, where the inputs are presented to the ANN

279 along with the targeted output. For each neuron in the hidden layers, the input vector (including a

280 unity element, the bias) is multiplied by a vector of weights using a scalar product. Although the

281 most commonly used learning algorithm in ANN is the backpropagation algorithm (fitted with

282 gradient descent and gradient descent with momentum), it is often time-consuming for a practical

283 point of view as it requires low learning rates for stable learning. Whereas algorithms such as

284 conjugate gradient, quasi-Newton, and Levenberg-Marquardt provide alternative ways which are

285 faster yet efficient. Two-hidden-layers have been thought as the most effective ANN architecture

286 ([Jones, 2004](#)), therefore, it is used in this study. For each input vector containing a combination of

287 SMOS T_{bs} , there is an associated target containing an SMD value. The output of the ANN is

288 compared with the desired value, and the weights are adjusted by minimising a cost function (i.e.,

289 MSE). The minimisation has been achieved by the Broyden–Fletcher–Goldfarb–Shanno (BFGS)

290 neural network training algorithm ([Fletcher, 2013](#)), and the conjugate gradient training algorithm

291 ([Bishop, 1995](#)). The BFGS algorithm is a variable metric or quasi-Newton method, where the

292 quadratic error function evaluated at w near to the minimum w^* is considered as the following

293 equation:

$$294 \quad E(w) = E(w^*) + \frac{1}{2}(w - w^*)^T H(w - w^*) \quad (11)$$

295 By differentiating Equation (11), the location of the minimum w^* can be calculated as:

$$296 \quad g \equiv \nabla E(w) = H(w - w^*) = 0 \quad (12)$$

297 The minimum w^* can therefore be calculated as:

298 $w^* = w - H^{-1}g$ (13)

299 where the vector $-H^{-1}g$ is the Newton direction and when validated at any w on a quadratic error
300 surface, it will direct to the minimum of the error function w^* .

301 For the conjugate gradient training algorithm, to achieve consecutive conjugate search directions,
302 the gradient $g \equiv \nabla E(w)$ of the error surface at the next point must be a minimum in the current
303 search direction d_j , which is achieved when:

304 $d_{j+1}Hd_j = 0$ (14)

305 where H is the Hessian matrix appraised at the point w_{j+1} . This direction search method is called
306 conjugate. Full mathematical descriptions of the two training algorithms used in this study can be
307 found in the aforementioned literature.

308 **4. Results**

309 In this study, four performance indicators are used: Pearson product moment correlation
310 coefficient (r), Mean squared error (MSE), Nash-Sutcliffe Efficiency (NSE) ([Nash and Sutcliffe,](#)
311 [1970](#)), and Root Mean Square Error (RMSE).

312 **4.1 Time series plots of XAJ SMD and SMOS soil moisture observations**

313 We have selected the days on which both the SMOS-BEC and the SMOS-CATDS have available
314 soil moisture data. This selection is to make a fair judgement between the two products because
315 during the same time period SMOS-CATDS has more available data than SMOS-BEC. The time
316 series plots of the XAJ SMD and the two soil moisture products are presented in Fig. 3. It can be
317 seen that the SMD demonstrates a high variability with seasons, with nadir (lower SMD indicates
318 wetter soil) often occurring in winter where evapotranspiration demand is the lowest. On the other

319 hand during the summer season, the hot temperature and increased evapotranspiration lead to an
320 overall drier surface soil (i.e., high SMD). For the two satellite soil moisture products, it is clear
321 to observe that they are slightly discriminated from each other. In order for a better visualisation,
322 two enlarged time series plots (i.e., during a winter period and a summer period, respectively) are
323 presented in Fig. 4. Interestingly during the winter period (Fig. 4a) when the soil is frozen, SMOS-
324 BEC observations are significantly less available than the SMOS-CATDS's. In order to appraise
325 the hydrological values of the SMOS-CATDS frozen soil moisture data, the correlation coefficient
326 is calculated against the XAJ SMD ($r = -0.76$). This high correlation value indicates that even
327 under the frozen condition, some of the satellite retrieved T_b s data are still useful for soil moisture
328 estimation. Due to the limited availability of the SMOS-BEC data during the winter season, its
329 correlation is not calculated here. During the summer period (Fig. 4b), data availability for both
330 products is higher than in the winter period, and their soil moisture values are closer to each other.
331 It can be seen from both products in the two enlarged plots that the summer soil (averagely around
332 $0.15 \text{ m}^3/\text{m}^3$) is generally drier than the winter soil (averagely around $0.25 \text{ m}^3/\text{m}^3$), which agrees
333 with the XAJ SMD fluctuations. The results of SMD estimation directly from the two SMOS soil
334 moisture products are presented in the later section of the paper.

335 **4.2 SMD estimation using SMOS brightness temperature as input**

336 **4.2.1 Input data selection**

337 As discussed in Section 3.2, an appropriate selection of incidence angles of the SMOS brightness
338 temperature observations is necessary to ensure the best SMD retrieval. In this study, data selection
339 is carried out by using a full embedding (embedding means a selection of inputs from all the
340 possible inputs) calculation with the gamma (Γ) from the GT as a metric. This approach tests every
341 combination of data inputs to determine which combination yields the smallest absolute gamma

342 value. If there are m scalar inputs then there are $2^m - 1$ possible embeddings (i.e., 16383
343 embeddings in this case). Although this method is more time consuming, it is more comprehensive.
344 The full embedding result is demonstrated by a histogram plot in Fig. 5, which shows the frequency
345 of embeddings with a given gamma statistic. It can be seen that the histogram tends to be a
346 Gaussian distribution, indicating that the choice of embedding is largely driven by statistical
347 variations in the data (Jones, 1998). The best inputs combination is from the embedding that gives
348 the lowest gamma value, which is the combination of H polarisation at the incidence angles of
349 32.50° , 37.50° , 47.50° , 52.50° , 57.50° and V polarisation at the incidence angles of 27.50° , 32.50° ,
350 37.50° , 42.50° , 57.50° . Although Fig. 2 shows that incidence angle 27.50° at H polarisation is more
351 correlated with the SMD, it may contain some duplicated features with other angles (called
352 redundancy) and is therefore excluded. There is a similar reason for those angles that also have a
353 high correlation with the SMD, but are not selected after the full embedding test. The gamma
354 statistic given by this combination is $\Gamma = 0.048$, and the gradient utilised to calculate the gamma
355 statistic is 0.51 (A) which roughly indicate that the output SMD is a relatively simple function of
356 the ten T_{bs} inputs. A model with low Γ and low A is considered to be the best scenario for
357 modelling. Therefore using the selected ten brightness temperature data solely should be efficient
358 in modelling the SMD variations.

359 The quantity of the training data to predict the desirable output is again analysed by the M-test,
360 which is useful in deciding whether there is sufficient data to provide an asymptotic gamma
361 estimate and subsequently a reliable model. The results of the M-test are presented in Fig. 6. To
362 select the most suitable training-data length, a trade-off between the best gamma and standard error
363 results, and the longest testing-data length is made. As a result, the 356 data length produces the
364 best trade-off result. The corresponding gamma and standard error are 0.061 and 0.0062,

365 respectively. The small values of both statistics illustrate that the gamma test is relatively accurate.
366 The results of the aforementioned tests give a clear image that it is possible to build a nonlinear
367 predictive model utilising 356 data points.

368 **4.2.2 SMD estimation using LLR model**

369 After selection of the input data, the LLR model is trained (between the 1st-356th data points) and
370 tested (between the 357th-434th data points) on the simulated SMD data from XAJ. It is important
371 to choose the optimal number of nearest neighbours (p_{\max}) in LLR so that the best model
372 performance can be achieved. This has been identified by the trial and error method. The procedure
373 is carried out by repeating the training and testing processes for another four times over different
374 training-testing data combinations using the 4-fold cross-validation (i.e., shifting the data by 108
375 each time) so that there is a total of five training-testing data combinations (including the training-
376 testing data combination obtained from the M-test). In this way, all the data are tested at least once
377 instead of just using the original testing data. The trial and error results (not normalised) are
378 presented in Table 1. It is observed that the MSE varies with different p_{\max} values and divided
379 groups, indicating that both factors are important in controlling the LLR modelling performance.
380 The generally low MSE values observed in group 1 clearly reveal the usefulness of the M-test. It
381 is still difficult to judge the most appropriate p_{\max} value based on those individual case results.
382 Therefore, it is necessary to average them so that a smooth trial and error curve can be obtained
383 (Fig. 7). The close MSE values between the testing and the training demonstrate that the LLR
384 model is quite stable in simulating the SMD values from the selected T_{bs} inputs. The LLR model
385 with p_{\max} at three generally gives the lowest MSE value and is therefore implemented hereafter.
386 The performance of the LLR technique is measured by three global statistics (NSE, r , and RMSE).
387 Fig. 8 shows the scatter plots of the LLR computed and the XAJ simulated SMD during the training

388 and testing periods. LLR shows a rather satisfactory performance (NSE = 0.88, $r = 0.94$, RMSE =
389 0.008 m) during the training phase in estimating the SMD. The majority of the data points are
390 saturated around the 45° line (dotted line) indicating that the model is well trained. Points far above
391 the bisector line signify over-estimation whereas points far below the dotted line mean under-
392 estimation. The training outcome illustrates the degree to which the LLR model explains SMD
393 variation as a function of the ten T_{bs} inputs, while the effectiveness of the model is judged during
394 the testing phase. It is clear to see that the LLR model performs very well during the testing phase
395 (NSE = 0.85, $r = 0.93$, RMSE = 0.009 m). A large number of saturated data points around the dotted
396 line signifies that there is a surprisingly excellent match between the modelled SMD and the XAJ
397 SMD. The used LLR algorithm has been double checked by disrupting the SMD target in the
398 testing datasets and changing the input file, and its performance remains the same. Therefore, it is
399 believed that LLR model is rather suitable for estimating SMD from T_{bs} . Fig. 9 shows that the
400 median of the XAJ simulated SMD is higher than the LLR modelled. Nevertheless, the LLR model
401 performs well for both low and high SMD values as the 5/25% and 75/95% percentiles of the XAJ
402 and the model estimated SMD match well.

403 **4.2.3 SMD estimation using ANN models**

404 The LLR model is then compared with two ANN models (i.e., the BFGS training algorithm ANN
405 and the conjugate gradient training algorithm ANN, respectively). The feedforward network used
406 in this work has two hidden layers. Various tests have been done to determine the optimal ANN
407 architecture. In the ANN conjugate gradient model, above 5 neurons in the hidden layer, the results
408 do not improve anymore, therefore 10-5-5-1 ANN structure is adopted. For the ANN BFGS model,
409 the feedforward 10-8-8-1 ANN is found to be the most suitable. The size of the sufficient training
410 dataset has been determined as 356 through the M-test, and the target MSE has been identified as

411 0.061 (normalised) to avoid the potential overtraining problem. Scatter plots of the two ANN
412 models during the training and testing phases are illustrated in Fig. 10, and their statistical
413 performances are indicated accordingly in the figure. It is seen in the statistics summary table
414 (Table 2), that the SMDs estimated by ANNs are inferior to the estimates by the LLR model for
415 both the training and testing parts. Box plots comparing the spread of the ANN estimated SMDs
416 with the XAJ simulated are also shown in Fig. 9. The plot indicates that both ANN models do not
417 capture the extreme low SMD values well (the 5% whiskers), but they perform acceptably in
418 estimating extreme high SMD values (the upper 95% whiskers). In addition, both ANN models
419 are comparatively poorer in modelling high SMD values (75% percentile) than LLR. The ANN-
420 BFGS is able to simulate low SMD well (25% percentile), while the ANN-conjugate shows less
421 capability in this aspect. On the other hand, the ANN-conjugate's simulation is able to produce the
422 closest mean SMD value to the XAJ's, while the ANN-BFGS's mean is more deviated. Generally,
423 the statistics results of the study indicate that the SMD predictive capability by the ANN-conjugate
424 is stronger than the ANN-BFGS.

425 **4.3 SMD estimation using SMOS soil moisture as input**

426 To further evaluate the proposed method, a comparison study is carried out to derive the SMD
427 directly from the two SMOS soil moisture products. LLR model is adopted for this purpose
428 because this is a mono-variable regression problem (i.e., to derive from one of the SMOS soil
429 moisture products into the SMD). If ANN is used it will have only one input node which makes
430 the ANN model ineffective. The quantities of the training and the testing data are again analysed
431 by the M-test. The M-test results show that the most suitable training data period for the SMOS-
432 BEC and the SMOS-CATDS is 1st-216th and 1st-220th, respectively, and the rest of the data are
433 used as the testing dataset. The optimal number of P_{\max} in LLR model is found to be 13 in both

434 data input cases. The SMD estimation results are illustrated in Fig. 11. The goodness of fit is
435 indicated by NSE, r , and RMSE statistics. The statistical performances between the two cases are
436 close to each other, indicating there is no significant difference between the two soil moisture
437 products. The poor results during both the training and the testing phases reveal that those soil
438 moisture products generated using the in-situ soil moisture networks and the numerical weather
439 modelling outputs as the evaluating target are not hydrologically suitable. Although both ANN
440 models are not capable of surpassing the LLR technique, their SMD estimations are still much
441 better than those derived from the SMOS soil moisture directly (as shown in Table 2). Therefore,
442 the proposed method using the SMOS multi-angle brightness temperatures is a more efficient way.

443 **5. Discussion and conclusions**

444 This paper describes a novel approach for the first time to estimate hydrological SMD directly
445 from the SMOS multi-angle brightness temperatures with both the H and V polarisations using
446 nonlinear modelling techniques. A well-proven gamma test is also employed to further improve
447 the input data feature selection process. The use of LLR and ANNs with the BFGS NN training
448 algorithm and the conjugate gradient training algorithm have been presented in this study. Both
449 the radial BFGS ANN training algorithm and the conjugate gradient training algorithm perform
450 well in estimating the SMD data, yet both fail to achieve the highest possible results. On the other
451 hand, the training and testing results demonstrate that the LLR model is surprisingly good at
452 capturing the interrelations between SMD and T_{bs} over ANNs. All the SMD values estimated from
453 the proposed nonlinear methods achieve outstanding accuracies compared with those derived from
454 the standard SMOS soil moisture products (both from the SMOS-BEC and the SMOS-CATDS).

455 The results from the LLR model are quite puzzling due to a large number of data points perfectly
456 matching with the predicted SMD values, in both the training phase and the testing phase. One

457 obvious suspicion is the model overfits the training data, however this has been excluded using the
458 combination of the training data and the testing data because an over-trained model cannot perform
459 well in the testing phase. Our explanation is such a phenomenon is caused by two nearby points
460 which have identical or almost the same SMD values. This happens if the distance between them
461 is very small, and is more likely to happen with LLR model which is local in comparison with
462 other global models such as ANN. A local model breaks the whole data points into local groups.
463 For a special case when $p_{\max}=1$, a value to be estimated at a certain point will be totally decided
464 by its nearest neighbour. If its nearest neighbour is close enough a zero error could be achieved.
465 However if the local data points are very sparse then its nearest neighbour will be quite far away,
466 and the estimated value will have a large error. This explains why there are so many points on the
467 perfectly matched line, while there are still many data points off it. The overall results indicate that
468 the LLR technique has a huge potential to provide hydrologists with valuable information on the
469 application of satellite brightness temperature for SMD estimation, which has not been explored
470 before. The current study could form the basis for efficient satellite data assimilation into real-time
471 flood forecasting systems. The LLR model evaluated in this paper is numerically very efficient
472 and is capable of retrieving SMD fast enough to be assimilated into such systems.

473 In this study the 'ground truth' is based on the SMD simulation from the XAJ model. One may
474 argue that a hydrological model's soil moisture state variable has no physical meaning and its
475 purpose is purely to facilitate a model's flow simulation, hence it has no direct connection with
476 the real-field soil moisture. Moreover, as Keith Beven states in [Beven \(2012\)](#) there are many
477 models with different parameter values which could produce equally good flow simulations (called
478 the equifinality effect) because those models are all optimised with the same flow simulation. As
479 a result, models with similar flow simulation accuracy could have very distinct values in their soil

480 moisture state variables. To explore this argument, we have carried out some numerical
481 experiments to demonstrate that although the absolute SMD values could vary greatly between
482 different model parameter sets, their response patterns to soil moisture changes are almost identical
483 because they are driven by the same precipitation and evapotranspiration processes with the
484 identical physical response mechanisms. Therefore, the SMD pattern is the true reflection of the
485 soil moisture changes in the real field, and this justifies the usage of SMD derived from the
486 hydrological model as the ‘ground truth’ for assessing soil moisture data quality. However, the
487 SMD and the real-field soil moisture represent different aspects of the soil moisture condition. A
488 regression formula is needed to convert the satellite observations into hydrological SMD as shown
489 in Fig. 8 and 10 (to derive hydrological SMD from the SMOS raw data using ANN and LLR) and
490 Fig. 11 (to convert from the SMOS soil moisture product into hydrological SMD). To make a fair
491 comparison, the regression formulas with the similar complexity are used in both cases.

492 The accuracy of the SMD estimation is largely dependent on the relationship of the training dataset
493 with the target output. The presence of erroneous values and under/over estimation in the training
494 dataset hampers the model performance. Although larger training data sizes generally yield better
495 results, it is challenging to decide what size is large enough, especially when the analysed data
496 period is short. At the moment, the rule of two-third data for training and one-third data for testing
497 is still popular albeit such a method lacks consideration of the data characteristics. In addition,
498 there is no commonly recognised method for input data feature selection and quality check, which
499 has hampered many modelling developments. This is because some input data sets carry duplicated
500 features (high redundancy), which can make the model over-complicated (over-fitting). Also, if
501 the inherent errors in the input data exceed the model’s capability, it is rather difficult for the model
502 to perform well, even the model itself is good enough. This study demonstrates the informative

503 capability of the GT and the M-test in the input data selection for nonlinear model constructions.
504 It is hoped that this approach could be generalised to benefit various research areas including
505 hydrology, meteorology and where input data feature selection is needed.

506 The mismatch between the satellite footprint and catchment scale is an important issue that should
507 be considered in the hydrological application of soil moisture products. In this study, the chosen
508 catchment has a compatible size with the satellite footprint, therefore the mismatch is not an issue
509 in this case. The effect of larger or smaller catchments should be explored in future studies. Since
510 the adopted LLR model is data based, the optimal model could change for various soil type,
511 catchment size, land cover and climate regions. The proposed scheme has to be applied to
512 individual catchments with their own model development for SMD estimation. With more studies
513 using the proposed method, it could be feasible to build a look-up table in which users can search
514 for the model structure and parameters so that it can be utilised in ungauged catchments as well.
515 Finally, it should be noted that the SMD produced from this paper cannot be directly used in
516 agricultural management or other disciplines because there is no universal soil moisture product
517 for all purposes. Nevertheless, for any specific application field, the proposed method can be easily
518 adopted to it by changing the targeted soil moisture (e.g., to change SMD to volumetric soil
519 moisture to be used in agriculture).

520 **Acknowledgments**

521 We acknowledge the U.S. Geological Survey for making available daily streamflow records
522 (<http://waterdata.usgs.gov/nwis/rt>). The NLDAS-2 data sets used in this article can be obtained
523 from the NASA Land Data Assimilation Systems website
524 (<http://ldas.gsfc.nasa.gov/nldas/NLDAS2forcing.php>), the SMOS level-3 brightness temperatures
525 are from the Centre Aval de Traitement des Données SMOS (CATDS; <http://www.catds.fr/>), and

526 the SMOS level-3 soil moisture datasets can be downloaded from the SMOS Barcelona Expert
527 Centre (BEC; <http://www.smos-bec.icm.csic.es/>) and the CATDS.

528

529

530

531 **References**

532 Ahmad, S., Kalra, A., Stephen, H., 2010. Estimating soil moisture using remote sensing data: A
533 machine learning approach. *Advances in Water Resources*, 33(1): 69-80.

534 Aires, F., Paul, M., Prigent, C., Rommen, B., Bouvet, M., 2011. Measure and exploitation of
535 multisensor and multiwavelength synergy for remote sensing: 2. Application to the
536 retrieval of atmospheric temperature and water vapor from MetOp. *Journal of Geophysical*
537 *Research: Atmospheres* (1984–2012), 116(D2).

538 Al-Yaari, A., Wigneron, J.-P., Ducharne, A., Kerr, Y., De Rosnay, P., De Jeu, R., Govind, A., Al
539 Bitar, A., Albergel, C., Munoz-Sabater, J., 2014. Global-scale evaluation of two satellite-
540 based passive microwave soil moisture datasets (SMOS and AMSR-E) with respect to
541 Land Data Assimilation System estimates. *Remote Sensing of Environment*, 149: 181-195.

542 Al - Shrafany, D., Rico - Ramirez, M.A., Han, D., Bray, M., 2013. Comparative assessment of
543 soil moisture estimation from land surface model and satellite remote sensing based on
544 catchment water balance. *Meteorological Applications*, 21(3): 521-534.

545 Al Bitar, A., Leroux, D., Kerr, Y.H., Merlin, O., Richaume, P., Sahoo, A., Wood, E.F., 2012.
546 Evaluation of SMOS soil moisture products over continental US using the

547 SCAN/SNOTEL network. *Geoscience and Remote Sensing, IEEE Transactions on*, 50(5):
548 1572-1586.

549 Bartholomé, E., Belward, A., 2005. GLC2000: a new approach to global land cover mapping from
550 Earth observation data. *International Journal of Remote Sensing*, 26(9): 1959-1977.

551 Beven, K.J., 2012. *Rainfall-runoff modelling: the primer*. John Wiley & Sons, Oxford, UK.

552 Bishop, C.M., 1995. *Neural networks for pattern recognition*. Oxford university press.

553 Brodzik, M.J., Knowles, K.W., 2002. EASE-Grid: A versatile set of equal-area projections and
554 grids. *Discrete global grids*, 5: 110-125.

555 Calvet, J.-C., Wigneron, J.-P., Walker, J., Karbou, F., Chanzy, A., Albergel, C., 2011. Sensitivity
556 of passive microwave observations to soil moisture and vegetation water content: L-band
557 to W-band. *Geoscience and Remote Sensing, IEEE Transactions on*, 49(4): 1190-1199.

558 Chen, X., Yang, T., Wang, X., Xu, C.-Y., Yu, Z., 2013. Uncertainty Intercomparison of Different
559 Hydrological Models in Simulating Extreme Flows. *Water resources management*, 27(5):
560 1393-1409.

561 Daly, C., Neilson, R.P., Phillips, D.L., 1994. A statistical-topographic model for mapping
562 climatological precipitation over mountainous terrain. *Journal of applied meteorology*,
563 33(2): 140-158.

564 Dehghani, M., Saghafian, B., Nasiri Saleh, F., Farokhnia, A., Noori, R., 2014. Uncertainty analysis
565 of streamflow drought forecast using artificial neural networks and Monte - Carlo
566 simulation. *International Journal of Climatology*, 34(4): 1169-1180.

567 Draper, C., Mahfouf, J.-F., Calvet, J.-C., Martin, E., Wagner, W., 2011. Assimilation of ASCAT
568 near-surface soil moisture into the SIM hydrological model over France. Hydrology and
569 Earth System Sciences, 15(12): 3829-3841.

570 Durrant, P.J., 2001. winGamma TM: a non-linear data analysis and modelling tool with
571 applications to flood prediction. PhD thesis, Cardiff University, P.O. Box 916, Cardiff,
572 CF24 3XF, Wales, UK.

573 Elsa, J., Ahmad, A.B., François, C., Arnaud, M., Philippe, R., Arnaud, Q., Lucie, B., 2013. CATDS
574 SMOS L3 soil moisture retrieval processor Algorithm Theoretical Baseline Document
575 (ATBD), [http://www.cesbio.ups-tlse.fr/SMOS_blog/wp-](http://www.cesbio.ups-tlse.fr/SMOS_blog/wp-content/uploads/2013/08/ATBD_L3_rev2_draft.pdf)
576 [content/uploads/2013/08/ATBD_L3_rev2_draft.pdf](http://www.cesbio.ups-tlse.fr/SMOS_blog/wp-content/uploads/2013/08/ATBD_L3_rev2_draft.pdf). Accessed on February 2, 2016.

577 Entekhabi, D., Njoku, E.G., O'Neill, P.E., Kellogg, K.H., Crow, W.T., Edelstein, W.N., Entin, J.K.,
578 Goodman, S.D., Jackson, T.J., Johnson, J., 2010. The soil moisture active passive (SMAP)
579 mission. Proceedings of the IEEE, 98(5): 704-716.

580 Evans, D., Jones, A.J., 2002. A proof of the Gamma test, Proceedings of the Royal Society of
581 London A: Mathematical, Physical and Engineering Sciences. The Royal Society, pp.
582 2759-2799.

583 Fletcher, R., 2013. Practical methods of optimization. John Wiley & Sons, West Sussex, England.

584 Gan, T.Y., Dlamini, E.M., Biftu, G.F., 1997. Effects of model complexity and structure, data
585 quality, and objective functions on hydrologic modeling. Journal of Hydrology, 192(1):
586 81-103.

587 Han, D., Kwong, T., Li, S., 2007. Uncertainties in real - time flood forecasting with neural
588 networks. *Hydrological processes*, 21(2): 223-228.

589 Han, E., Merwade, V., Heathman, G.C., 2012. Implementation of surface soil moisture data
590 assimilation with watershed scale distributed hydrological model. *Journal of hydrology*,
591 416: 98-117.

592 Hansen, M., R. DeFries, J.R.G. Townshend, and R. Sohlberg, 1998. UMD Global Land Cover
593 Classification. In: 1 Kilometer, Department of Geography, University of Maryland,
594 College Park, Maryland, 1981-1994 (Ed.).

595 Hornik, K., Stinchcombe, M., White, H., 1989. Multilayer feedforward networks are universal
596 approximators. *Neural networks*, 2(5): 359-366.

597 Ireland, G., Volpi, M., Petropoulos, G.P., 2015. Examining the Capability of Supervised Machine
598 Learning Classifiers in Extracting Flooded Areas from Landsat TM Imagery: A Case Study
599 from a Mediterranean Flood. *Remote Sensing*, 7(3): 3372-3399.

600 Islam, T., Rico-Ramirez, M.A., Han, D., Srivastava, P.K., 2012. Artificial intelligence techniques
601 for clutter identification with polarimetric radar signatures. *Atmospheric Research*, 109:
602 95-113.

603 Jaafar, W.W., Han, D., 2011. Variable selection using the gamma test forward and backward
604 selections. *Journal of Hydrologic Engineering*, 17(1): 182-190.

605 Jacquette, E., Al Bitar, A., Mialon, A., Kerr, Y., Quesney, A., Cabot, F., Richaume, P., 2010.
606 SMOS CATDS level 3 global products over land, *Remote Sensing*. International Society
607 for Optics and Photonics, pp. 78240K-78240K-6.

608 Jones, A., 1998. The WinGamma User Guide. Copyright: University of Wales, Cardiff, 2001.

609 Jones, A.J., 2004. New tools in non-linear modelling and prediction. *Computational Management*
610 *Science*, 1(2): 109-149.

611 Kerr, Y.H., Waldteufel, P., Richaume, P., Wigneron, J.P., Ferrazzoli, P., Mahmoodi, A., Al Bitar,
612 A., Cabot, F., Gruhier, C., Juglea, S.E., 2012. The SMOS soil moisture retrieval algorithm.
613 *Geoscience and Remote Sensing, IEEE Transactions on*, 50(5): 1384-1403.

614 Kerr, Y.H., Waldteufel, P., Wigneron, J.-P., Delwart, S., Cabot, F., Boutin, J., Escorihuela, M.-J.,
615 Font, J., Reul, N., Gruhier, C., 2010. The smos mission: New tool for monitoring key
616 elements of the global water cycle. *Proceedings of the IEEE*, 98(5): 666-687.

617 Kerr, Y.H., Waldteufel, P., Wigneron, J.-P., Martinuzzi, J., Font, J., Berger, M., 2001. Soil
618 moisture retrieval from space: The Soil Moisture and Ocean Salinity (SMOS) mission.
619 *Geoscience and Remote Sensing, IEEE Transactions on*, 39(8): 1729-1735.

620 Koncar, N., 1997. *Optimisation methodologies for direct inverse neurocontrol*, University of
621 London.

622 Liu, X., Zhao, D., Xiong, R., Ma, S., Gao, W., Sun, H., 2011. Image interpolation via regularized
623 local linear regression. *Image Processing, IEEE Transactions on*, 20(12): 3455-3469.

624 Massari, C., Brocca, L., Moramarco, T., Trambly, Y., Lescot, J.-F.D., 2014. Potential of soil
625 moisture observations in flood modelling: Estimating initial conditions and correcting
626 rainfall. *Advances in Water Resources*, 74: 44-53.

627 McMullan, K., Brown, M., Martín-Neira, M., Rits, W., Ekholm, S., Marti, J., Lemanczyk, J., 2008.
628 SMOS: The payload. *Geoscience and Remote Sensing, IEEE Transactions on*, 46(3): 594-
629 605.

630 Mitchell, K.E., Lohmann, D., Houser, P.R., Wood, E.F., Schaake, J.C., Robock, A., Cosgrove,
631 B.A., Sheffield, J., Duan, Q., Luo, L., 2004. The multi - institution North American Land
632 Data Assimilation System (NLDAS): Utilizing multiple GCIP products and partners in a
633 continental distributed hydrological modeling system. *Journal of Geophysical Research:*
634 *Atmospheres* (1984–2012), 109(D7). DOI:10.1029/2003JD003823

635 Nash, J., Sutcliffe, J., 1970. River flow forecasting through conceptual models part I—A
636 discussion of principles. *Journal of Hydrology*, 10(3): 282-290.

637 Noori, R., Karbassi, A., Moghaddamia, A., Han, D., Zokaei-Ashtiani, M., Farokhnia, A.,
638 Gousheh, M.G., 2011. Assessment of input variables determination on the SVM model
639 performance using PCA, Gamma test, and forward selection techniques for monthly stream
640 flow prediction. *Journal of Hydrology*, 401(3): 177-189.

641 Peel, M.C., Finlayson, B.L., McMahon, T.A., 2007. Updated world map of the Köppen-Geiger
642 climate classification. *Hydrology and Earth System Sciences Discussions*, 4(2): 439-473.

643 Penrose, R., 1955. A generalized inverse for matrices, *Mathematical proceedings of the Cambridge*
644 *philosophical society*. Cambridge Univ Press, pp. 406-413.

645 Penrose, R., 1956. On best approximate solutions of linear matrix equations, *Mathematical*
646 *Proceedings of the Cambridge Philosophical Society*. Cambridge Univ Press, pp. 17-19.

647 Pi, H., Peterson, C., 1994. Finding the embedding dimension and variable dependencies in time
648 series. *Neural Computation*, 6(3): 509-520.

649 Pinson, P., Nielsen, H.A., Madsen, H., Nielsen, T.S., 2008. Local linear regression with adaptive
650 orthogonal fitting for the wind power application. *Statistics and Computing*, 18(1): 59-71.

651 Remesan, R., Shamim, M., Han, D., 2008. Model data selection using gamma test for daily solar
652 radiation estimation. *Hydrological processes*, 22(21): 4301-4309.

653 Rodríguez-Fernández, N.J., Aires, F., Richaume, P., Kerr, Y.H., Prigent, C., Kolassa, J., Cabot, F.,
654 Jiménez, C., Mahmoodi, A., Drusch, M., 2015. Soil Moisture Retrieval Using Neural
655 Networks: Application to SMOS. *IEEE TRANSACTIONS ON GEOSCIENCE AND
656 REMOTE SENSING (TGRS)*, 53(11).

657 Rodriguez-Fernandez, N., Richaume, P., Aires, F., Prigent, C., Kerr, Y., Kolassa, J., Jimenez, C.,
658 Cabot, F., Mahmoodi, A., 2014. Soil moisture retrieval from SMOS observations using
659 neural networks, *Geoscience and Remote Sensing Symposium (IGARSS)*, 2014 IEEE
660 International. *IEEE*, pp. 2431-2434.

661 Shi, P., Chen, C., Srinivasan, R., Zhang, X., Cai, T., Fang, X., Qu, S., Chen, X., Li, Q., 2011.
662 Evaluating the SWAT model for hydrological modeling in the Xixian watershed and a
663 comparison with the XAJ model. *Water resources management*, 25(10): 2595-2612.

664 SMOS-BEC, 2015. SMOS Barcelona Expert Centre ocean and land products description,
665 <http://cp34-bec.cmima.csic.es/doc/BEC-SMOS-0001-PD.pdf>. Accessed on February 2,
666 2016.

667 Srivastava, P.K., Han, D., Ramirez, M.R., Islam, T., 2013a. Machine learning techniques for
668 downscaling SMOS satellite soil moisture using MODIS land surface temperature for
669 hydrological application. *Water resources management*, 27(8): 3127-3144.

670 Srivastava, P.K., Han, D., Rico-Ramirez, M.A., Al-Shrafany, D., Islam, T., 2013b. Data fusion
671 techniques for improving soil moisture deficit using SMOS satellite and WRF-NOAH land
672 surface model. *Water resources management*, 27(15): 5069-5087.

673 Srivastava, P.K., Han, D., Rico Ramirez, M.A., Islam, T., 2013c. Appraisal of SMOS soil moisture
674 at a catchment scale in a temperate maritime climate. *Journal of Hydrology*, 498: 292-304.

675 Stefánsson, A., Končar, N., Jones, A.J., 1997. A note on the gamma test. *Neural Computing &*
676 *Applications*, 5(3): 131-133.

677 Sun, H., Liu, H., Xiao, H., He, R., Ran, B., 2003. Use of local linear regression model for short-
678 term traffic forecasting. *Transportation Research Record: Journal of the Transportation*
679 *Research Board*(1836): 143-150.

680 Tehrany, M.S., Pradhan, B., Jebur, M.N., 2014. Flood susceptibility mapping using a novel
681 ensemble weights-of-evidence and support vector machine models in GIS. *Journal of*
682 *Hydrology*, 512: 332-343.

683 Trambly, Y., Bouaicha, R., Brocca, L., Dorigo, W., Bouvier, C., Camici, S., Servat, E., 2012.
684 Estimation of antecedent wetness conditions for flood modelling in northern Morocco.
685 *Hydrology and Earth System Sciences*, 16(11): 4375-4386.

686 Tsui, A.P., Jones, A.J., De Oliveira, A.G., 2002. The construction of smooth models using irregular
687 embeddings determined by a gamma test analysis. *Neural Computing & Applications*,
688 10(4): 318-329.

689 USDA, 2005. Global Soil Regions Map. United States Department of Agriculture, Natural
690 Resources Conservation Service Soils,
691 http://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/use/worldsoils/?cid=nrcs142p2_05
692 [4013](#). Accessed on February 2, 2016.

693 Walker, J.P., Willgoose, G.R., Kalma, J.D., 2004. In situ measurement of soil moisture: a
694 comparison of techniques. *Journal of Hydrology*, 293(1): 85-99.

695 Wang, L., Qu, J.J., 2009. Satellite remote sensing applications for surface soil moisture monitoring:
696 A review. *Frontiers of Earth Science in China*, 3(2): 237-247.

697 Wei, E.-B., Liu, S.-B., Wang, Z.-Z., Tong, X.-L., Dong, S., Li, B., Liu, J.-Y., 2014. Emissivity
698 measurements of foam-covered water surface at l-band for low water temperatures. *Remote*
699 *Sensing*, 6(11): 10913-10930.

700 Wigneron, J.-P., Kerr, Y., Waldteufel, P., Saleh, K., Escorihuela, M.-J., Richaume, P., Ferrazzoli,
701 P., De Rosnay, P., Gurney, R., Calvet, J.-C., 2007. L-band Microwave Emission of the
702 Biosphere (L-MEB) Model: Description and calibration against experimental data sets over
703 crop fields. *Remote Sensing of Environment*, 107(4): 639-655.

704 Xia, Y., Mitchell, K., Ek, M., Sheffield, J., Cosgrove, B., Wood, E., Luo, L., Alonge, C., Wei, H.,
705 Meng, J., 2012. Continental - scale water and energy flux analysis and validation for the
706 North American Land Data Assimilation System project phase 2 (NLDAS - 2): 1.

707 Intercomparison and application of model products. *Journal of Geophysical Research:*
708 *Atmospheres* (1984–2012), 117(D3).

709 Zhao, R., 1980. The Xinanjiang model, . *Hydrological Forecasting Proceedings Oxford*
710 *Symposium, IASH 129*, 351-356.

711 Zhao, R., 1992. The Xinanjiang model applied in China. *Journal of Hydrology*, 135(1): 371-381.

712 Zhao, R., Liu, X., 1995. The Xinanjiang model. In: Singh, V. (Ed.), *Computer models of watershed*
713 *hydrology*. Water Resource Publications, Highlands Ranch, pp. 215-232.

714 Zhuo, L., Dai, Q., Han, D., 2015a. Evaluation of SMOS soil moisture retrievals over the central
715 United States for hydro-meteorological application. *Physics and Chemistry of the Earth,*
716 *Parts A/B/C*, 83-84: 146–155.

717 Zhuo, L., Dai, Q., Han, D., 2015b. Meta - analysis of flow modeling performances—to build a
718 matching system between catchment complexity and model types. *Hydrological Processes,*
719 *29(11): 2463–2477.*

720 Zhuo, L., Dai, Q., Islam, T., Han, D., 2016. Error distribution modelling of satellite soil moisture
721 measurements for hydrological applications. *Hydrological Processes.*
722 DOI:10.1002/hyp.10789

723 Zhuo, L., Han, D., 2016a. Could operational hydrological models be made compatible with
724 satellite soil moisture observations? *Hydrological Processes.* DOI:10.1002/hyp.10804.

725 Zhuo, L., Han, D., 2016b. Misrepresentation and amendment of soil moisture in conceptual
726 hydrological modelling. *Journal of Hydrology*, 535: 637-651.

727 Zhuo, L., Han, D., Dai, Q., Islam, T., Srivastava, P.K., 2015c. Appraisal of NLDAS-2 Multi-Model
728 Simulated Soil Moistures for Hydrological Modelling. *Water Resources Management*,
729 29(10): 3503-3517.

730 Zurada, J.M., 1992. *Introduction to artificial neural systems*. West St. Paul, Minnesota, USA.

731

732

Table 1. Trial and error results of finding the best number of nearest neighbours (p_{max}) in the LLR model.

p_{max}	group 1		group 2		group 3		group 4		group 5		mean	
	training	testing	training	testing	training	testing	training	testing	training	testing	training	testing
1	8.1E-05	1.3E-04	1.4E-04	1.4E-04	1.0E-04	8.4E-05	8.4E-05	1.1E-04	1.1E-04	7.3E-05	1.0E-04	1.1E-04
2	7.6E-05	1.1E-04	1.2E-04	1.4E-04	6.4E-05	9.0E-05	6.8E-05	5.6E-05	8.9E-05	7.0E-05	8.3E-05	9.3E-05
3	7.1E-05	7.6E-05	1.0E-04	1.1E-04	6.0E-05	9.2E-05	6.9E-05	1.0E-04	1.0E-04	6.7E-05	8.1E-05	9.0E-05
4	6.1E-05	8.6E-05	9.6E-05	1.1E-04	7.5E-05	1.1E-04	7.8E-05	1.4E-04	1.1E-04	6.4E-05	8.3E-05	1.0E-04
5	7.3E-05	7.9E-05	1.1E-04	1.0E-04	1.1E-04	1.1E-04	9.3E-05	1.6E-04	1.2E-04	8.8E-05	1.0E-04	1.1E-04
6	6.7E-05	1.0E-04	1.2E-04	1.4E-04	1.4E-04	6.8E-05	1.3E-04	2.0E-04	1.2E-04	9.2E-05	1.2E-04	1.2E-04

(Note: The performance is measured by the mean squared error (MSE is in the unit of m^2). The datasets (i.e., 434 in total) have been divided into five groups so that all of them can be tested at least once. Group 1 comprises the training data of 1-356, and testing data of 357-434 from the M-test; group 2 comprises the training data of 1-326, and testing data of 327-434; group 3 comprises the training data of 109-434, and testing data of 1-108; group 4 comprises the training data of 1-107, 216-434, and testing data of 108-215; group 5 comprises the training data of 1-216, 326-434, and testing data of 217-325. The mean MSE results are used to determine the optimal p_{max} value in LLR model.)

Table 2. Summary of the model performances.

	Training			Testing		
	<i>NSE</i>	<i>r</i>	<i>RMSE(m)</i>	<i>NSE</i>	<i>r</i>	<i>RMSE(m)</i>
LLR	0.88	0.94	8.0E-3	0.85	0.93	9.0E-3
ANN- conjugate	0.74	0.86	1.2E-2	0.64	0.81	1.4E-2
ANN- BFGS	0.77	0.88	1.2E-2	0.60	0.79	1.4E-2
SMOS-BEC	0.55	0.74	1.5E-2	0.34	0.60	1.8E-2
SMOS-CATDS	0.53	0.73	1.5E-2	0.35	0.61	1.8E-2

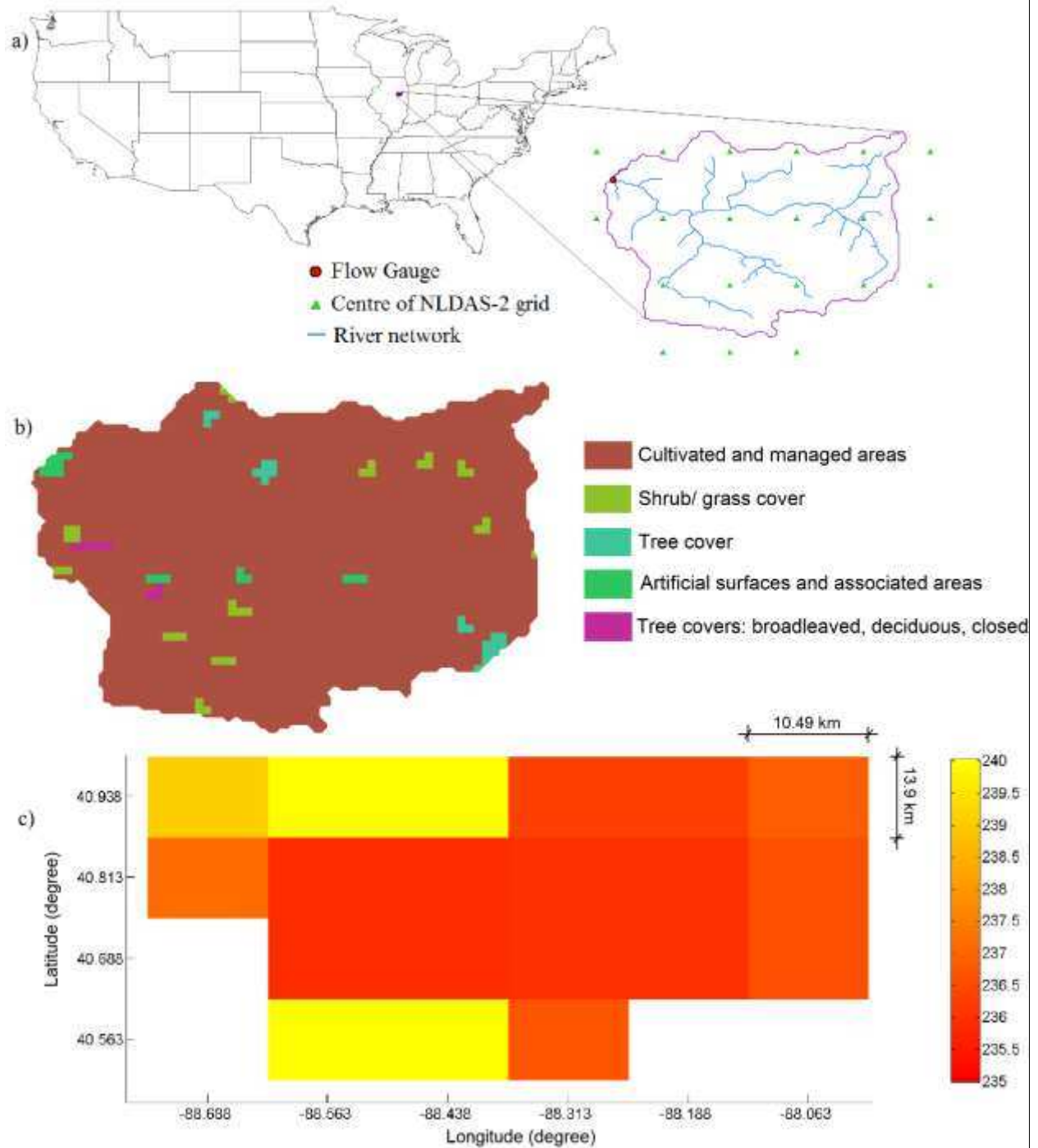


Figure 1. a) Geographical location of the study area with river network, flow gauge and NLDAS-2 grids; b) GIS extracted GLC2000 land-use map; c) spatial variations of the retrieved SMOS brightness temperature (in kelvins) data on 13/01/2010 at the ascending overpass, with the H polarisation and incidence angle of 32.5° for the catchment area (it has been transformed into NLDAS-2 grids at $0.125^\circ \times 0.125^\circ$ grid spacing for easier analysis).

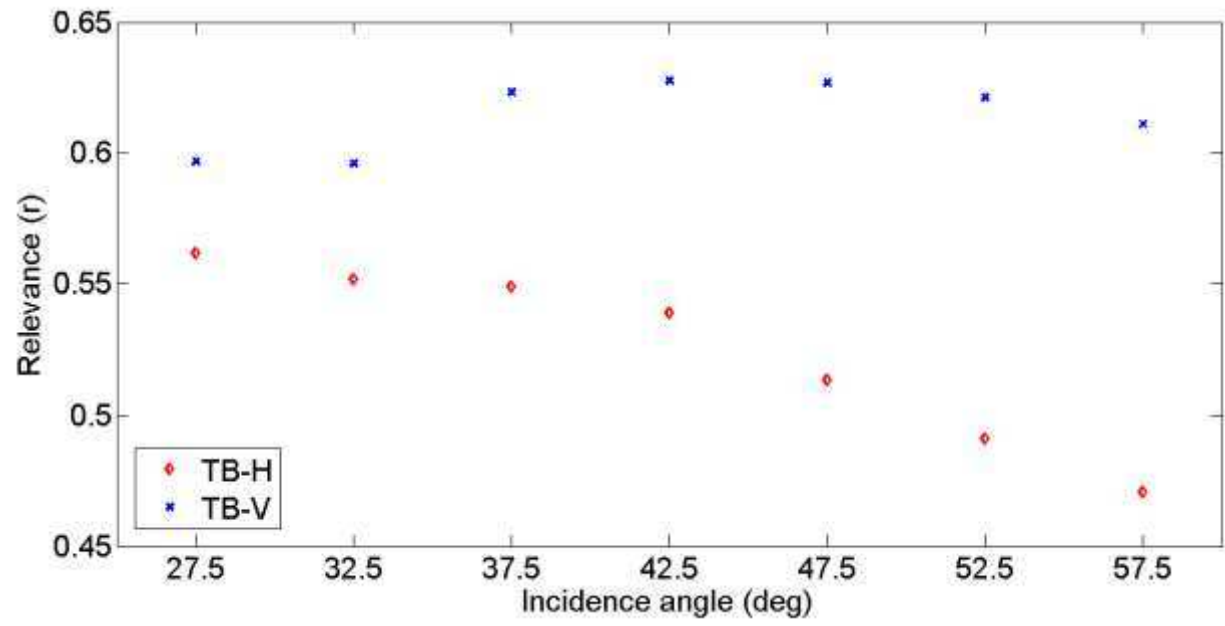


Figure 2. Correlations r between the SMOS multangular brightness temperatures with H and V polarisations and the XAJ SMD.

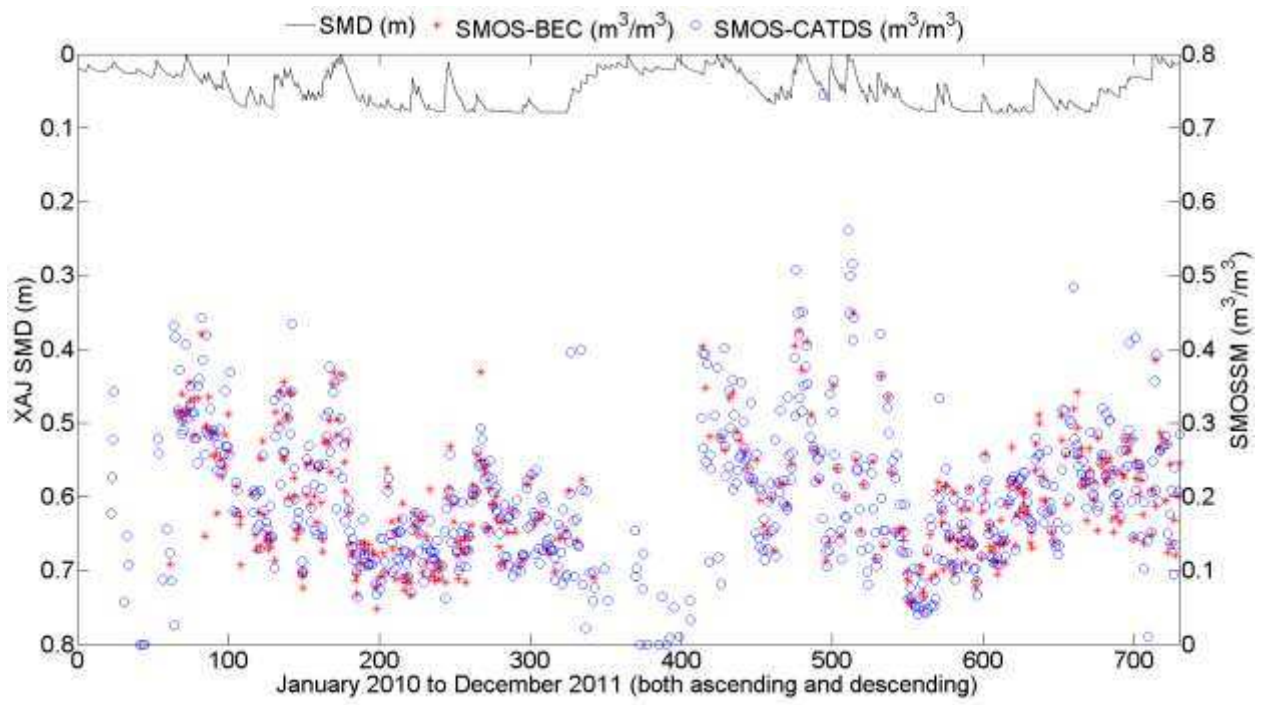


Figure 3. Time series plots of the XAJ SMD and the two SMOS soil moisture products (indicated as SMOSSM in the y-axis label) from CATDS and BEC, respectively.

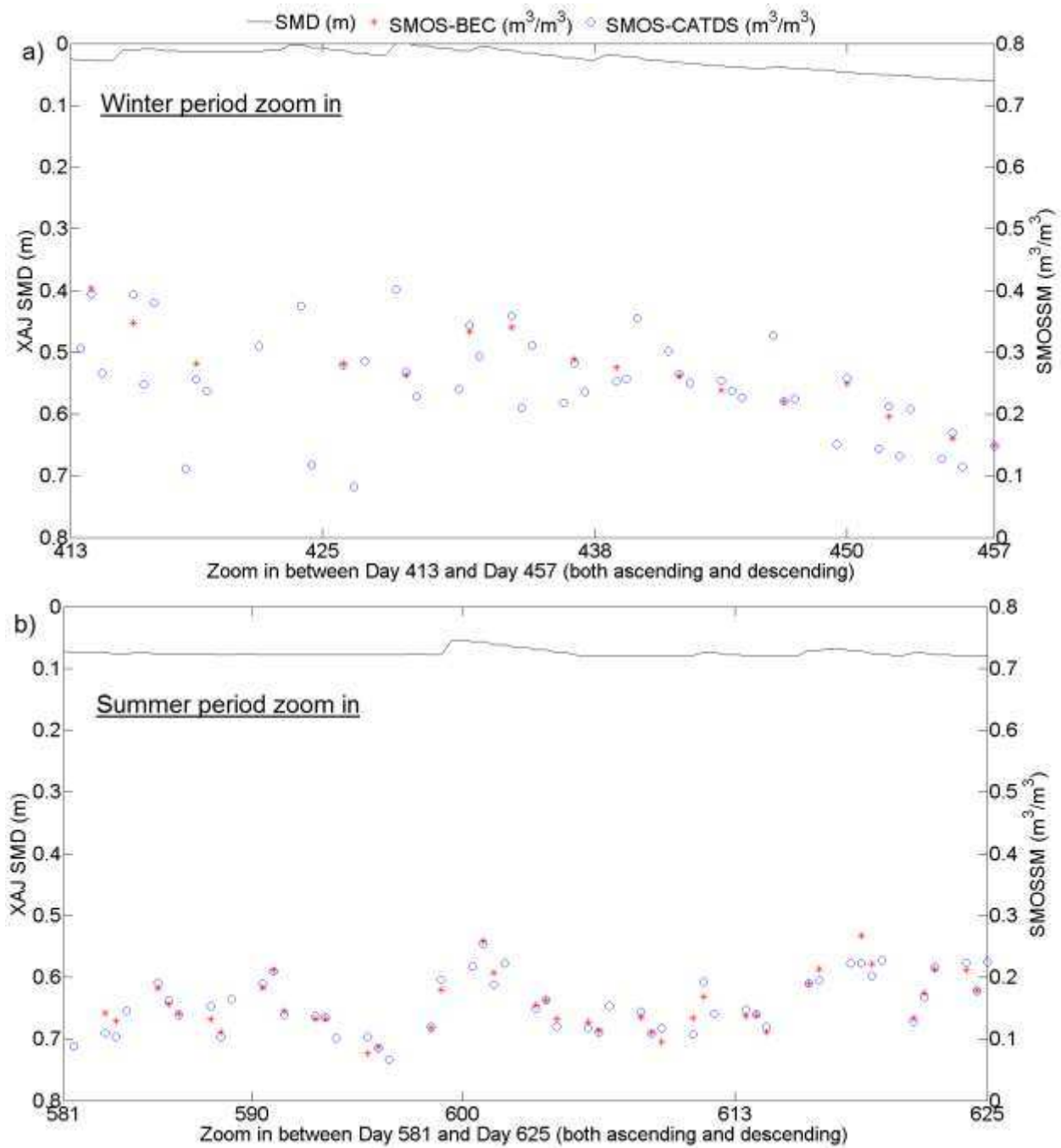


Figure 4. Two enlarged time series plots of the XAJ SMD and the two SMOS soil moisture products (indicated as SMOSSM in the y-axis label) from CATDS and BEC, respectively: a) between Day 413 and Day 457 (a winter period), and b) between Day 581 and Day 625 (a summer period).

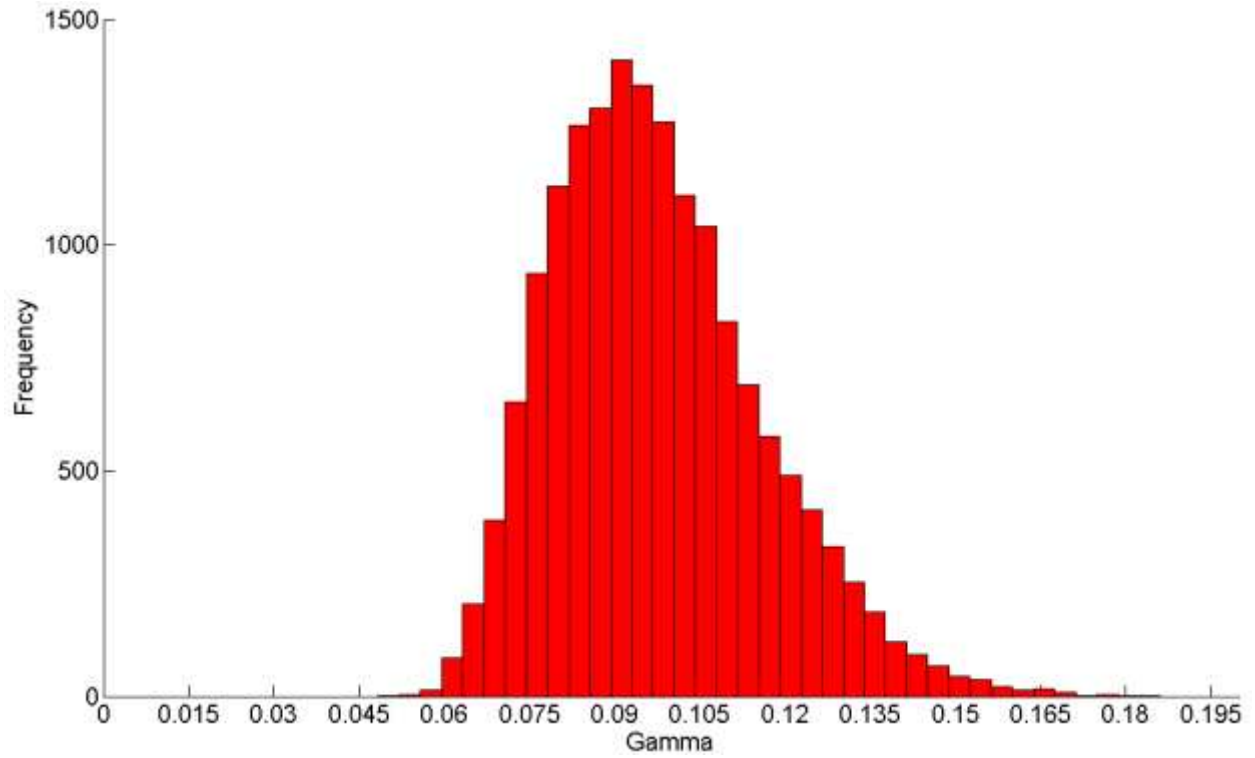


Figure 5. The histogram of the full embedding calculation, with the gamma (Γ) from the gamma test as a metric.

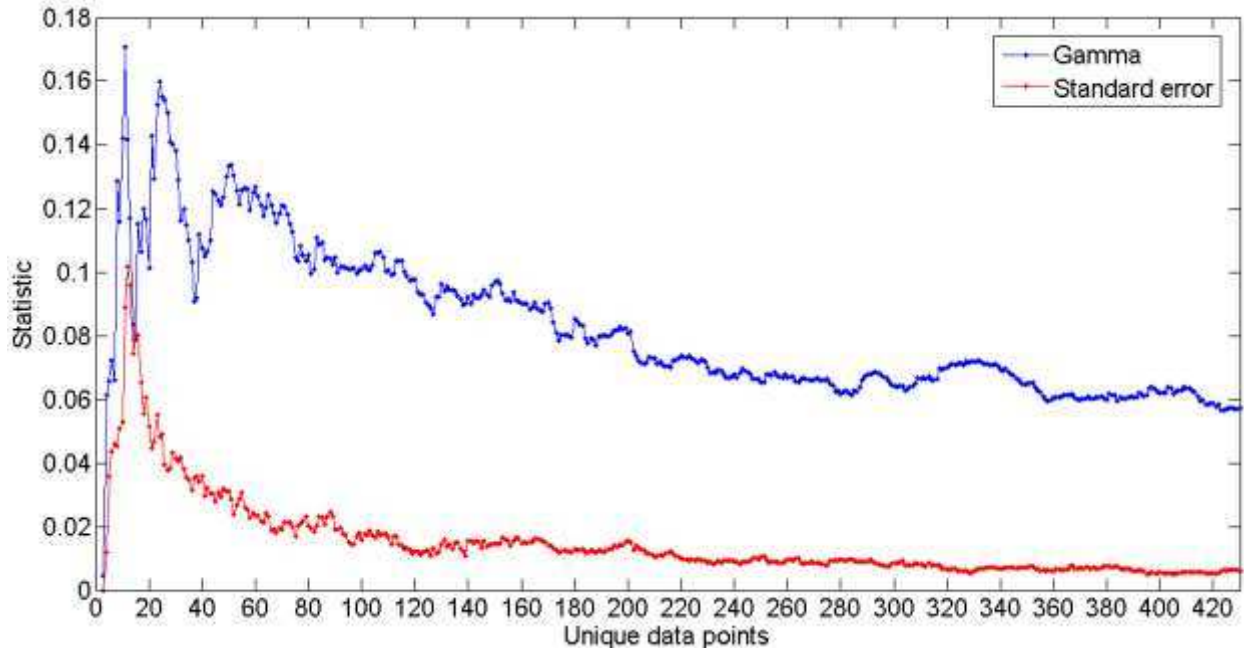


Figure 6. *M*-test results. It indicates an asymptotic convergence of the gamma (Γ) to a value of 0.061 at 356 data length, and the corresponding standard error at the convergent point is 0.0062.

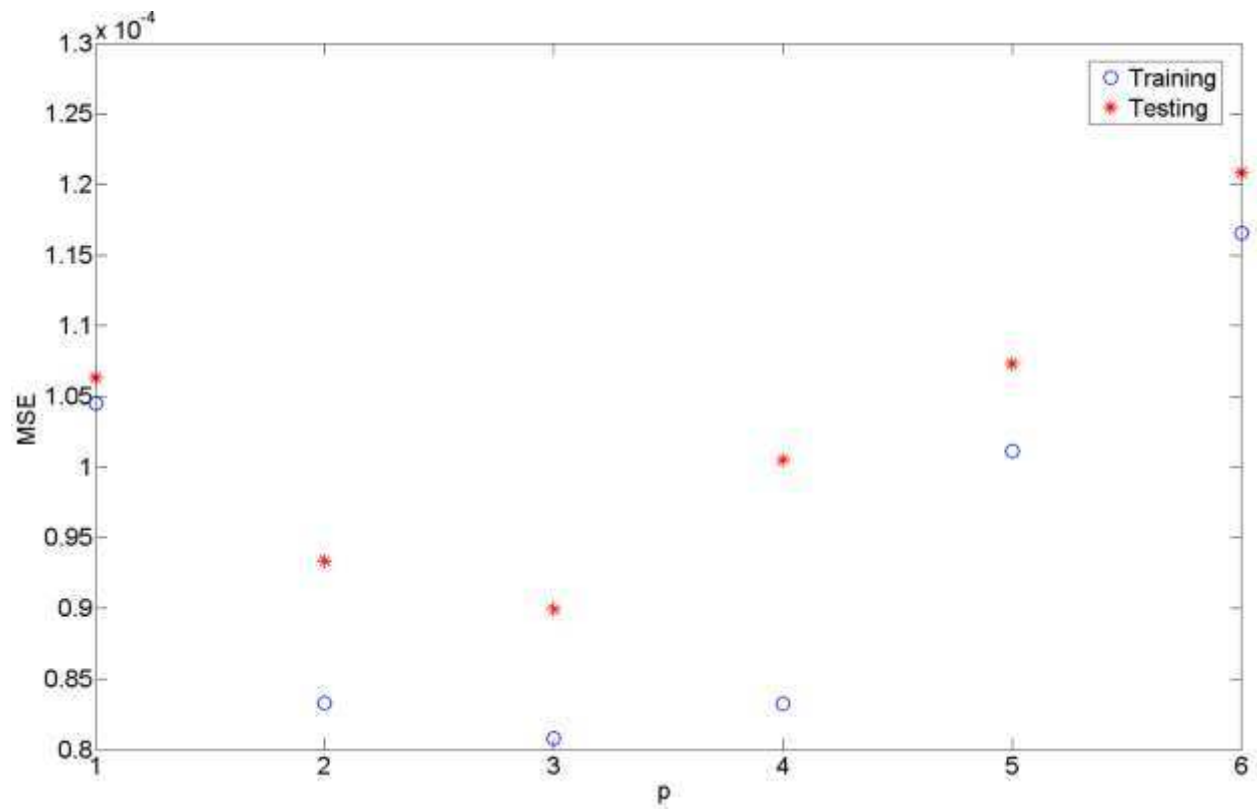


Figure 7. Trial and error result to find the optimal p_{max} value in the LLR modelling.

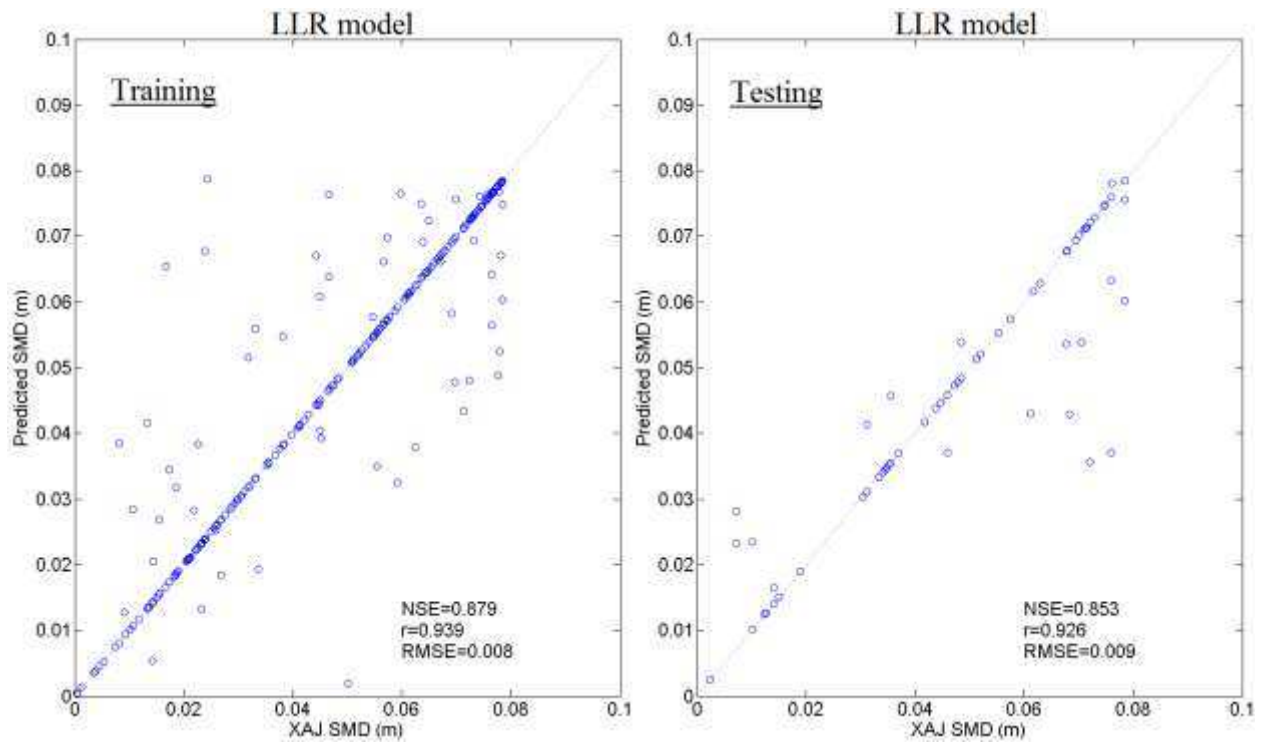


Figure 8. SMD simulated by the LLR model. It shows the scatter plots of the LLR computed and the XAJ simulated SMD during the training and testing periods. It is noted that *RMSE* is in the unit of metre.

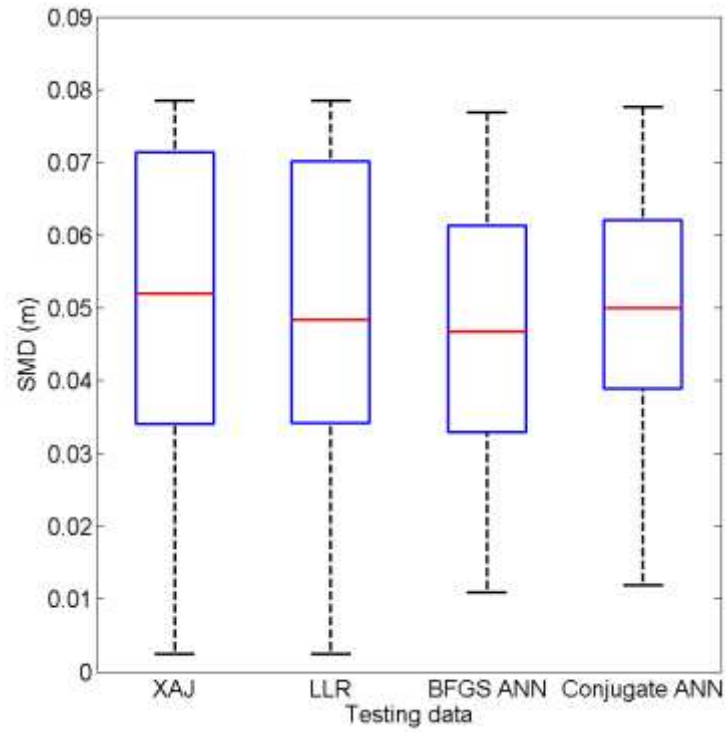


Figure 9. The statistical plot of the XAJ simulated SMD and the models estimated SMD during the testing phase. The boxes indicate 25–75% percentiles. The whiskers extend from 5% to 95% percentile values. The red line represents the median value of the data.

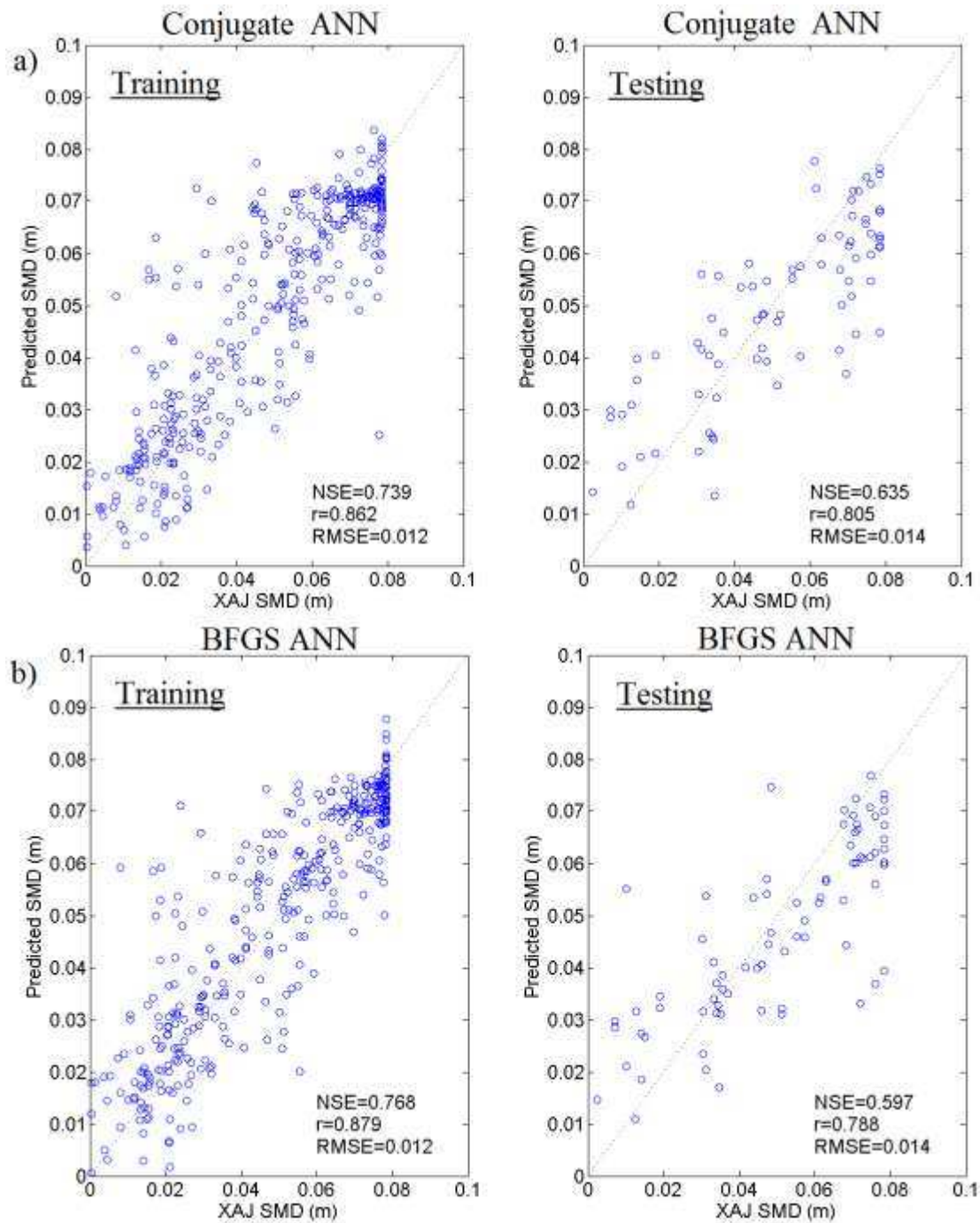


Figure 10. SMD simulated by the ANN models. a) shows the scatter plots of the conjugate ANN computed and the XAJ simulated SMD during the training and testing periods; b) presents the scatter plots of the BFGS ANN computed and the XAJ simulated SMD during the training and testing periods. It is noted that *RMSE* is in the unit of metre.

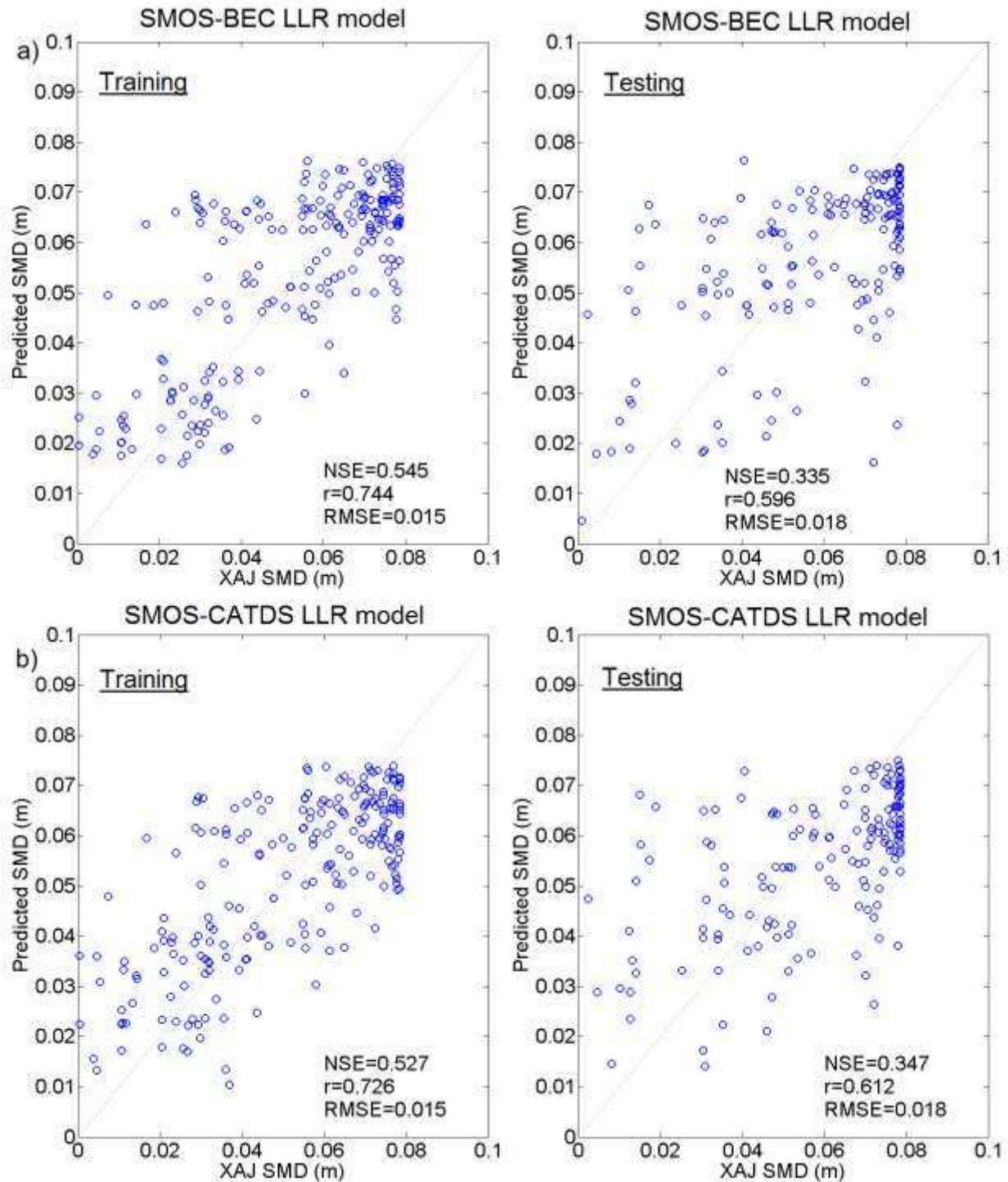


Figure 11. SMD estimation using LLR model and SMOS soil moisture input: a) from SMOS-BEC; b) from SMOS-CATDS. It is noted that *RMSE* is in the unit of metre.