1   **Multi-source hydrological soil moisture state estimation using data fusion optimisation**

2

3   Lu Zhuo[1*], Dawei Han[1]

4   [1]WEMRC, Department of Civil Engineering, University of Bristol, Bristol, UK

5   [*]Correspondence: lu.zhuo@bristol.ac.uk

6   **Abstract**

7   Reliable estimation of hydrological soil moisture state is of critical importance in operational

8   hydrology to improve the flood prediction and hydrological cycle description. Although there

9   have been a number of soil moisture products, they cannot be directly used in hydrological

10   modelling. This paper attempts for the first time to build a soil moisture product directly

11   applicable to hydrology using multiple data sources retrieved from SAC-SMA (soil moisture),

12   MODIS (land surface temperature), and SMOS (multi-angle brightness temperatures in H-V

13   polarizations). The simple yet effective Local Linear Regression model is applied for the data

14   fusion purpose in the Pontiac catchment. Four schemes according to temporal availabilities of

15   the data sources are developed, which are pre-assessed and best selected by using the well-

16   proven feature selection algorithm Gamma Test. The hydrological accuracy of the produced

17   soil moisture data is evaluated against the Xinanjiang hydrological model's soil moisture

18   deficit simulation. The result shows that a superior performance is obtained from the scheme

19   with the data inputs from all sources ($NSE = 0.912$, $r = 0.960$, $RMSE = 0.007$ m). Additionally

20   the final daily-available hydrological soil moisture product significantly increases the Nash-

21   Sutcliffe efficiency by almost 50 % in comparison with the two most popular soil moisture

22   products. The proposed method could be easily applied to other catchments and fields with

23   high confidence. The misconception between the hydrological soil moisture state variable and

24   the real-world soil moisture content, and the potential to build a global routine hydrological

25   soil moisture product are discussed.

30   **1. Introduction**

31   Soil moisture is a key element in the hydrological cycle, regulating evapotranspiration,

32   precipitation infiltration and overland flow (Wanders et al., 2014). For hydrological

33   applications, the antecedent wetness condition of a catchment is among the most significant

34   factors for accurate flow generation processes (Berthet et al., 2009; Matgen et al., 2012a).

35   (Norbiato et al., 2008) reported that initial wetness conditions are essential for efficient flash

36   flood alerts. Additionally an operational system requires reliable hydrological soil moisture

37   state updates to reduce the time drift problem (Aubert et al., 2003; Berg and Mulroy, 2006;

38   Dumedah and Coulibaly, 2013). However, currently there is no available soil moisture product

39   that can be used directly in hydrology modelling, primarily because soil moisture is difficult to

40   define and there is no single shared meaning in various disciplines (Romano, 2014).

41   Although there have been many soil moisture measuring projects (e.g., satellite missions such

42   as Advanced Scatterometer (ASCAT), Soil Moisture and Ocean Salinity (SMOS), and Soil

43 Moisture Active Passive (SMAP); ground-based networks such as Soil Climate Analysis

44 Network (SCAN), U.S. Surface Climate Observing Reference Networks (USCRN), and

45 COsmic-ray Soil Moisture Observing System (COSMOS)), they are not sufficiently used in

46 hydrology due to the following reasons: 1) misconception between the hydrological soil

47 moisture state variable and the real-field soil moisture content (Zhuo and Han, 2016a); 2)

48 unawareness of data availability and strength/weakness of different data sources; 3) the existing

49 soil moisture products are mainly evaluated against point-based ground soil moisture

50 observations or airborne retrievals which have significant spatial mismatch (both horizontally

51 and vertically) to catchment-scales, and are therefore less applicable to hydrological modelling

52 (Pierdicca et al., 2013); 4) underutilisation of multiple data sources (e.g., multi-angle raw

53 observations by satellite sensors).

54 Some studies have attempted to directly utilise the existing soil moisture products (i.e., data

55 from satellites, land surface models, and in-situ methods directly) for flood prediction

56 improvement, for example (Brocca et al., 2010) explored that utilising the soil water index

57 from ASCAT sensor could improve runoff prediction mainly if the initial catchment wetness

58 conditions were unknown; (Aubert et al., 2003) assimilated in-situ soil moisture observations

59 into a simple rainfall-runoff model and acquired better flow prediction performance ; (Javelle

60 et al., 2010) suggested that estimations of antecedent soil moisture conditions were useful in

61 improving flash flood forecasts at ungauged catchments; contrarily (Chen et al., 2011)'s study

62 showed assimilating ground-based soil moisture observations was generally unsuccessful in

63 enhancing flow prediction; and (Matgen et al., 2012b) revealed that satellite soil moisture

64 products added little or no extra value for hydrological modelling. Clearly those results are

65   rather mixed. Challenges remain in integrating soil moisture estimated outside the hydrological

66   field into hydrological models. We believe if a hydrologically directly applicable soil moisture

67   product could be produced, the aforementioned studies' results would be significantly

68   improved.

69   Therefore the aims of this paper are to clarify the aforementioned misconception between the

70   hydrological model's soil moisture state and the real-world soil moisture, assess the data

71   availabilities for direct hydrological soil moisture state estimation, and fuse those available

72   data sources using a hydrologically relevant approach. It is hoped that the final product has a

73   superior hydrological compatibility over the existing soil moisture products. To achieve these

74   aims, the Xinanjiang (XAJ) (Zhao, 1992) operational rainfall-runoff model is used as a target

75   to simulate flow and soil moisture state information (i.e., soil moisture deficit (SMD)) for the

76   Pontiac catchment in the central United States (U.S.). The reason for adopting XAJ is explained

77   in the following section. For the purpose of hydrological soil moisture state estimation, it is

78   effective to adopt the data driven method, which can map multiple data sources into the desired

79   dataset without computational burden. In this study the Local Linear Regression (LLR) model

80   is used. The multiple data sources applied in this study include the SMOS (Kerr et al., 2010b)

81   multi-angle brightness temperatures (T_bs) with both horizontal (H) and vertical (V)

82   polarizations, the Moderate Resolution Imaging Spectroradiometre (MODIS) (Wan, 2008)

83   land surface temperature, and the soil moisture product by SAC-SMA (Xia et al., 2014). The

84   detail explanations of those datasets are covered in the methodology section. A well-proven

85   feature selection algorithm Gamma Test (GT) (Stefánsson et al., 1997; Zhuo et al., 2016b) is

86   employed to pre-assess the selected data inputs and find the optimal combination of them for

4

87 soil moisture state calculation. In addition, an *M*-test (Remesan et al., 2008) is adopted to

88 explore the best size of the training data. The desired soil moisture product is trained and tested

89 by the XAJ SMD simulation. In total four data-input schemes are developed according to the

90 temporal availability of the selected data inputs, which are then combined to give a daily

91 hydrological soil moisture product.

92 Compared with previous work, our study contains the following new elements: i) a

93 hydrologically directly usable soil moisture product is proposed; ii) the GT and LLR techniques

94 are used for the first time in a data fusion of multiple data sources for hydrological soil moisture

95 state estimation; iii) the use of multiple data sources is useful, which allows data users to

96 analyse the availability of the different products and compare the relative benefits of them.

97 **2. Material and Methods**

98 **2.1 Study Area**

99 In this study, the Pontiac catchment (1,500 km^2, Figure 1) is used for the calibration and the

100 validation of the XAJ model. Pontiac (40.878°N, 88.636°W) lies on the north-flowing

101 Vermilion River, which is a tributary of the Illinois River of the state of Illinois, U.S. The worst

102 flood in this area occurred on December 4, 1982, cresting at 5.84 m above mean sea level

103 (MSL); and the most recent flood occurred on January 9, 2008, cresting at 5.75 m MSL, so this

104 catchment is likely located within a winter-flooding region. Pontiac is covered with moderate

105 canopy (the annual mean Normalized Difference Vegetation Index retrieved from the MODIS

106 satellite is around 0.4), when compared with a densely vegetated catchment, it has more

107 accurate soil moisture estimations from satellites (Al-Bitar et al., 2012). Based on the Köppen-

108 Geiger climate classification, this medium sized catchment is dominated mainly by hot summer

109 continental climate (Peel et al., 2007). With reference to the University of Maryland

110 Department Global Land Cover Classification, it is used primarily for agriculture purpose

111 (Bartholomé and Belward, 2005; Hansen, 1998). The soil mostly consists of Mollisols, which

112 has deep and high organic matter, and the nutrient-enriched surface soil is typically between

113 60-80 cm in depth (Webb et al., 2000). The study period is from January 2010 to December

114 2011. The reason for using this two-year period of data is due to the discontinuity of the flow

115 records in this catchment, and the selected period provides the most complete flow observations.

116 The North American Land Data Assimilation System 2 (NLDAS-2) (Mitchell et al., 2004)

117 provides precipitation and potential evapotranspiration information to run the XAJ model. Both

118 data forces are at $0.125°$ spatial resolution and have been converted to daily temporal resolution.

119 In order to use those distributed forcing into the lumped XAJ model, both forcing have been

120 interpolated with the area-weighted average method instead of the more complicated Kriging

121 approach, because the latter could produce errors if not well controlled (Wanders et al., 2014).

122 The average annual rainfall depth is about 954 mm, and the average annual potential

123 evapotranspiration is approximately 1670 mm. The daily observed flow data are acquired from

124 the U.S. Geological Survey.

125 **2.2 Hydrological Model**

126 The XAJ hydrological model is used for the simulation of SMD and river flow at a daily time

127 step. It is a simple lumped rainfall-runoff model with many applications performed in world-

128 wide catchments (Chen et al., 2013; Gan et al., 1997; Shi et al., 2011; Zhao, 1992; Zhao and

129   Liu, 1995; Zhuo et al., 2016a; Zhuo et al., 2015b). Since XAJ can obtain rather effective flow

130   modelling performances and require only two meteorological forcing (precipitation and

131   potential evapotranspiration) inputs (Peng et al., 2002), it is used more widely than the more

132   complicated semi-distributed/ fully-distributed hydrological models for operational

133   applications.

134   As shown in Figure 2, the XAJ model has three main components: evapotranspiration, runoff

135   generation, and runoff routing. XAJ consists of soil layers (upper, lower and deep) in its

136   evapotranspiration calculations. Because XAJ adopts the multi-bucket variable-size method in

137   its modelling concept, it has unfixed soil depths which is more effective than the fixed depths

138   models (Beven, 2012). Other widely used models such as PDM (Moore, 2007), VIC (Liang et

139   al., 1994), and ARNO (Todini, 1996) also follow this concept.

140   In XAJ, the three-layer soil moisture state variables are all calculated as SMD, which is an

141   important soil wetness variable in hydrology. SMD is defined as the amount of water to be

142   added to a soil profile to bring it to the field capacity (Calder et al., 1983; Rushton et al., 2006).

143   In this study, only the surface SMD (i.e., top layer) referring to the vegetation and the very thin

144   topsoil, is utilised as a hydrological soil moisture target. This is because the water held in the

145   top few centimetres of the soil has been widely recognised as a key variable associated with

146   water fluxes (Eltahir, 1998; Entekhabi and Rodriguez-Iturbe, 1994). Moreover the current

147   satellite technology is only capable of acquiring the Earth information from the outermost layer

148   of the soil. Therefore as a case study based on the XAJ model, we only focus on the surface

149   soil moisture state investigation here. Future research will focus on the root-zone soil moisture

150   product development by using a similar method proposed in this study.

151   In this study, a modified version of the XAJ model is adopted, and interested readers are

152   referred to (Zhuo and Han, 2016b) for more details. All the XAJ's 17 parameters are used

153   during the model calibration, which are shown in Table 1. In this study, the genetic algorithm

154   (Wang, 1991) is used for parameter optimisation. Based on the genetic algorithm result, minor

155   trial and error adjustments to the parameters EX, B, WUM, WLM and WDM are also carried out

156   to obtain the best model performance (Chen and Adams, 2006). The calibration and the

157   validation results (during January 2010-April 2011 and May 2011 to December 2011,

158   respectively) of the XAJ model can be found in (Zhuo et al., 2015a). Discussion regarding the

159   river flow and SMD simulation results in this catchment have been published in (Zhuo and

160   Han, 2016b), with Nash-Sutcliffe Efficiency (NSE) obtained larger than 0.80 during both the

161   calibration and validation periods. The results are not repeated here.

162   **2.3 Multiple Data Sources for Hydrological Soil Moisture State Estimation**

163   Data sources from SMOS, MODIS and SAC-SMA are used (Table 2). All data sources have

164   been converted into catchment-scale datasets by the area-weighted average method. The detail

165   description of each data source is given as follows. The main reason for choosing those three

166   data sources is due to their Near-Real-Time (NRT) availabilities (MODAPS Services, 2015;

167   Rodell, 2016) (SMOS becomes available in NRT recently (ESA Earth Online, 2016)), which

168   allows fast implementation in flood forecasting.

169   **2.3.1 SMOS Multi-angle Brightness Temperatures (SMOS-T$_b$s)**

170   The SMOS (1.4 GHz, L-band) Level-3 T_bs data covering the studying period are available from

171   the Centre Aval de Traitement des Données SMOS (CATDS) (Jacquette et al., 2010). The

172   reason for choosing the SMOS satellite is because compare with other satellite techniques (i.e.,

173   optical, and thermal infrared), microwave bands (especially with longer wavelength such as L-

174   band (21 cm)) can penetrate deeper into the soil (~ 5 cm) and have less interruptions from

175   weather conditions (Njoku and Kong, 1977). Additionally SMOS has a relatively longer period

176   of data record compares with other satellite missions such as SMAP. SMOS retrieves the

177   thermal emission from the Earth in both H and V polarizations with a wide ranges of incidence

178   angles from 0° to 60°. The observation depth of SMOS is approximately 5 cm with a spatial

179   resolution of 35-50 km depending on the incident angle and the deviation from the satellite

180   ground track (Kerr et al., 2012; Kerr et al., 2010a; 2001).

181   SMOS provides T_bs retrievals at all incidence angles averaged in 5° -width angle bins, which

182   have been transformed into the ground polarization reference frame (i.e., H, and V

183   polarizations). Therefore the number of the SMOS-T_bs inputs for the hydrological soil moisture

184   estimation can be as high as 24 (12 angle bins per polarization), with the centre of the first

185   angle bin at 2.5^o in both polarizations (Rodriguez-Fernandez et al., 2014). As satellite

186   progresses, any given location on the Earth's surface is scanned a number of times at various

187   incidence angles, depending on the location with respect to the satellite subtrack: the further

188   away, the fewer the angular acquisitions (Kerr et al., 2010b). The data availabilities of the

189   SMOS-T_bs are illustrated in Figure 3 (the availabilities for H and V polarizations are the same).

190   It can be seen that the data availabilities among various incidence angles are rather different.

191   In this study the only angle range that gives the most available record of data is from 27.5^o to

192   57.5^o (i.e., 7 for H and 7 for V polarization), which is therefore chosen for the hydrological soil

193   moisture development. This angle range is in line with the angle selection in (Rodriguez-

194   Fernandez et al., 2014). In addition the SMOS Level-3 soil moisture product from the CATDS

195   (SMOS-SM) is also acquired for a comparison with the estimated soil moisture product.

196   Retrievals that are potentially contaminated with Radio Frequency Interference have been

197   removed. Readers are referred to (Kerr et al., 2012) for a full description of the SMOS

198   retrieving algorithms, and (Njoku and Entekhabi, 1996) for a good knowledge of how passive

199   microwave relates to soil moisture variations.

200   **2.3.2 MODIS Land Surface Temperature (MODIS-LST**)

201   The MODIS/Terra (Earth Observing System AM-1 platform) (Wan, 2008) daily MOD11C1-

202   V5 land surface temperature covering the studying period is downloaded from the Land

203   Processes Distributed Active Archive Centre website. MODIS is chosen among other

204   operational optical satellites for its suitable features, mostly, due to its frequent revisiting time

205   and free NRT data availability. It measures 36 spectral bands between 0.405 and 14.385 μm,

206   and acquires data at three spatial resolutions 250 m, 500 m, and 1,000 m respectively while the

207   adopted MOD11C1 V5 product incorporates 0.05^o (5.6 km) spatial resolution. The benefit of

208   adding land surface temperature information is that previous studies have shown the variations

209   in soil moisture have a strong linkage with land surface temperature (Carlson, 2007; Goward

210   et al., 2002; Mallick et al., 2009). One reason is the changes of land surface temperature are

211   mainly affected by albedo and diurnal heat capacity, and the diurnal heat capacity is mainly

212   controlled by soil moisture (Price, 1980). (Wan, 2008) compared MOD11C1-V5 land surface

213   temperatures in 47 clear-sky cases with in situ measurement and revealed that the accuracy was

214   better than 1 K in the range from −10° to 58 °C in about 39 cases. Cloud-contaminated data

215   have been removed by a double-screening method, and its detail can be found in (Wan et al.,

216   2002).

217   **2.3.3 SAC-SMA Soil Moisture Estimation (SAC-SMA-SM)**

218   The reason for choosing the SAC-SMA land surface modelled soil moisture product is because

219   satellite can often have missing data due to various weather and canopy conditions (e.g., rainfall,

220   frozen weather, and vegetation coverage), so this daily dataset is essential in producing a

221   temporally completed hydrological soil moisture product. In this study, the surface soil

222   moisture (0-10 cm) simulated from the SAC-SMA model is selected. This is because its

223   estimated soil moisture gives a high accuracy against the observational soil moisture and a

224   good correlation with the XAJ SMD (Zhuo et al., 2015b). The daily SAC-SMA-SM is given

225   in   a   spatial   resolution   of   0.125°.   The   dataset   can   be   download   from

226   (http://www.emc.ncep.noaa.gov/mmb/nldas/). Readers are referred to (Xia et al., 2012) for a

227   full description of the SAC-SMA data products.

228   **2.3.4 Data Availabilities**

229   As shown in Table 2, the availability of the three data sources is rather different. Unlike SMOS

230   and MODIS, SAC-SMA-2 SM is a model based product which runs in a NRT mode, so it

231   produces valid data every day during the whole studying period. Whereas the two satellites'

232   data are more exiguous depends on weather and surface conditions. Compared with MODIS,

233   the SMOS's retrieval is even sparse and the biggest data shortage normally occurs in the winter

234 season where its returned microwave signal is mostly affected by frozen soils (Zhuo et al.,

235 2015a). Based on the data availability analysis, the proposed hydrological soil moisture product

236 is built from four data-input schemes as presented in Table 3. Those four schemes enable us to

237 test and compare the estimated soil moisture state more comprehensively. Since the continuity

238 of a soil moisture product is essential for any operational applications, SAC-SMA-SM is

239 included in all of the schemes.

240 **2.4 Data Fusion**

241 **2.4.1   Gamma Test (GT) for Feature Selection**

242 Before model building, it is important to carry out a feature selection process, because it can

243 simplify the model inputs, shorter training times, and reduce overfitting problems. In this study

244 a proper combination of the incidence angles from the SMOS T_bs is vital for the best soil

245 moisture state calculation. For this purpose, a feature selection method called GT is adopted. It

246 has been effectively used in numerous studies for model inputs selection (Durrant, 2001; Jaafar

247 and Han, 2011; Noori et al., 2011; Remesan et al., 2008; Tsui et al., 2002; Zhuo et al., 2016b).

248 In addition to the feature selection, GT can also give useful indication about the underlying

249 model complexity. It is a near-neighbour data analysis routine which determines the minimum

250 mean-squared error (*MSE*) that can be achieved based on the input-output dataset utilising any

251 continuous nonlinear models (Zhuo et al., 2016b). The calculated minimum *MSE* is referred as

252 the Gamma statistics and denoted as Γ. For detailed calculations about the GT algorithm,

253 interested readers are referred to (Koncar, 1997; Pi and Peterson, 1994; Stefánsson et al., 1997).

254 Here only the basic knowledge about the GT is shown:

255     $\{(x_i, y_i), \ 1 \le i \le M\}$ (1)

256   here the inputs $x_i \in R^m$ are vectors restricted by a closed bounded set $C \in R^m$, and their

257   corresponding outputs $y_i \in R$ are scalars, M stands for the sample points. The outputs y are

258   determined by the input vectors x that carry predictively useful messages. The only assumption

259   made is that their latent relationship is from the following function:

260     $y = f(x_1 \ldots x_m) + r$ (2)

261   here f is built up as a smooth model with r representing random noise. Without loss of generality,

262   the assumption of r noise distribution is that its mean is always zero, because all the constant

263   bias has been considered within the f model. Additionally r's variance ($Var(r)$) is restricted

264   within a set boundary. The observations' potential model is now defined within the class of

265   smooth functions.

266   The Γ is related to $N[i,k]$, which represents as the kth ($1 \le k \le p$) nearest neighbours of each

267   vector x_i ($1 \le i \le M$), written as $x_{N[i,k]}(1 \le k \le p)$, where p is a fixed integer. In order to

268   determine the Gamma function from the input vectors, the *Delta* function is used:

269     $\delta_M(k) = \dfrac{1}{M} \sum\limits_{i=1}^{M} \left| x_{N[i,k]} - x_i \right|^2$ ($1 \le k \le p$)   (3)

270   here the function $\left| x_{N[i,k]} - x_i \right|$ calculates the Euclidean distance. The Gamma function for its

271   output values is expressed as in Eq. 4, and the Γ can be determined from Eq. 3 and 4:

272     $\gamma_M(k) = \dfrac{1}{2M} \sum\limits_{i=1}^{M} \left| y_{N[i,k]} - y_i \right|^2$ ($1 \le k \le p$)   (4)

273   here $y_{N[i,k]}$ is the corresponding output values for the kth nearest neighbours x_i ($x_{N[i,k]}$). To

274   find Γ a least-squared regression line for the p points ($\delta_M(k)$, $\gamma_M(k)$) is built using the

275   following equation:

276    $\gamma = A\delta + \Gamma$ (5)

277   where Γ can be determined when δ is set as zero. The detailed explanation is:

278    $\gamma_M(k) \rightarrow Var(r)$, when $\delta_M(k) \rightarrow 0$ (6)

279   Eq. 5 gives us valuable information about the underlying system: not only that the Γ is a useful

280   indicator of the optimal MSE result that any smooth functions can achieve, but its gradient A

281   also provides guidance about the underlying model complexity (i.e., the steeper the gradient

282   the more sophisticated the model should be adopted). In this study, the winGamma^{TM} software

283   is used for GT calculation (Durrant, 2001). The mathematical feasibility of GT has been

284   published in (Evans and Jones, 2002).

2.4.2   *M*-test for Training Data Size Selection

286   A common practice in nonlinear modelling is to split the dataset into training and testing parts.

287   However there is no universal solution on how to divide the datasets (i.e., the proportion of

288   each part) so that the best modelling results could be obtained. Here, an M-test is carried out,

289   where M stands for the training data size. M-test is accomplished by calculating the Γ for

290   increasing the M value (i.e., expanding the training data) and exploring the resultant graph to

291   judge whether the Γ approaches a stable asymptote. Such an approach is straightforward and

292   effective in finding the optimal sizes of training and testing datasets, while avoiding overfitting

293   problems and reducing unsystematic attempts.

294   **2.4.3   Local Linear Regression (LLR)**

295   Various data fusion techniques have been developed (Prakash et al., 2012; Srivastava et al.,

296   2013; Wagner et al., 2012), however their methods require high computational time to run and

297   this, in a real-time flood forecasting framework, could not match the operational needs.

298   Comparatively, LLR model is a simpler method and requires relatively low computational time.

299   Therefore it is chosen in order to test if a simple method is able to provide effective

300   performance. LLR is a nonparametric regression model that has been applied in (Liu et al.,

301   2011; Pinson et al., 2008; Sun et al., 2003; Zhuo et al., 2016b) for forecasting and smoothing

302   purposes. LLR builds local linear regression based on the nearest points (p_{max}) of a targeted

303   point, and repeats such a process over the whole training dataset to produce a piecewise linear

304   model. There are many methodologies in selecting the p_{max}, in this study a method called

305   influence statistics is used (Durrant, 2001; Remesan et al., 2008), which is outlined as below.

306   Assume there are p_{max} nearest points, then the Eq. 7 can be built:

307     $Xm = y$ (7)

308   here X is a $p_{\max} \times d$ matrix which shows the d dimensional information of p_{max}, x_i are the

309   nearest points confined between 1 and p_{max}, y is the output vector with p_{max} dimension, and m

310   is a set of parameters formed in a vector, which plays an important role in mapping the solution

311   from X to y. Therefore Eq. 7 can be expanded as

15

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

(8)

In order to solve the equation, the following two conditions are set: a) if X is square and non-singular then Eq. (7) can be simply calculated as $m = X^{-1}y$; b) if X is not square or singular, Eq. (7) needs to be rearranged and m can be get by finding the minimum of:

$$|Xm - y|^2$$

(9)

with the distinct solution of:

$$m = X^{\#}y$$

(10)

where $X^{\#}$ is the pseudo-inverse matrix of X (Penrose, 1955; Penrose, 1956).

3. Results

In this section, different combinations of input data (Table 3) are adopted to examine their impacts on hydrological soil moisture estimation. XAJ SMD is used as a hydrological soil moisture state benchmark for the training and testing. More discussion about the misconception between the hydrological model's soil moisture state variable and the real-world soil moisture content is covered in Section 4. During GT and M-test processes, all data inputs need to be normalised so that their mean is zero and standard deviation is 0.5. This step is necessary in reducing the impacts of numerical difference from various inputs, hence improves the GT efficiency (Remesan et al., 2008). Five statistical indicators are used for the soil moisture estimation analysis: Pearson product moment correlation coefficient (r), MSE which is the

330   same value as the Gamma statistic Γ, Standard error (*SE*), *NSE* (Nash and Sutcliffe, 1970), and

331   Root Mean Square Error (*RMSE*).

3.1 Scheme 1: SMD Estimation Using SAC-SMA-SM as input

333   Although in this scheme, there is no need for data feature selection because only one data input

334   is involved, the GT is still carried out to explore the useful information about the underlying

335   relationship between the XAJ SMD and the SAC-SMA-SM. The calculated Gamma statistics

336   are shown in Table 4. The Γ of 0.072 indicates that the optimal *MSE* achievable using any

337   modelling technique is 0.072; and the small value of *SE* means the precision and accuracy of

338   the GT result. Γ is a significant target value in the *M*-test to find the most suitable training data

339   size. As presented in Figure 4a, when more training data (i.e., *M* increases in steps of one) is

340   used the Γ changes dramatically. Eventually at $M = 292$, Γ starts to stabilise around 0.072. The

341   *M*-test allows us to confidently apply the first 292 datasets to build a model of a given quality,

342   in the sense of predicting with a *MSE* around the asymptotic level. The corresponding Gamma

343   gradient (*A*) suggests the complexity of the underlying system: the larger the *A* value is the

344   more complex the system is. For example if *A* is significantly large, a more complicated model

345   like a Support Vector Machine might be required, but $A = 1.353$ in Scheme 1 is small (Remesan

346   et al., 2008), therefore a LLR model should be able to simulate the system. For LLR modelling,

347   its complexity level is controlled by the p_{max} parameter. As illustrated in Figure 5, p_{max} is

348   identified from a trial and error method. The procedure is by increasing the LLR p_{max} value

349   from 2 to 100 to analyse the variations of their corresponding Γ results. It can be seen from

350   Figure 5 that the smallest Γ is achieved at $p_{max} = 4$, which is therefore adopted for the LLR

351 modelling. The training and testing scatter plots for the LLR modelling are shown in Figure 6a.

352 It is observed that there are some points lying far above the bisector line during the training

353 period signifies higher estimations whereas some points sit far below the bisector line during

354 the testing period indicates under-estimation of the SMD. For the testing results, when XAJ

355 simulated soil moistures state have already reach the total dryness (i.e., XAJ SMD peaks at

356 around 0.080 m) the predicted soil moisture state is still in the drying progress. Figure 7a plots

357 the time series of the estimated and the targeted SMD. The plot shows that the estimated SMD

358 follows the seasonal trend of the soil moisture fluctuations well, so it is wetter during the winter

359 season and exsiccated during the hot summer season. However it is clear to see that the model

360 is not able to capture the extreme situations very well, especially during the wet season when

361 the XAJ SMD becomes smaller (e.g., between Day 300 and Day 350).

362 **3.2 Scheme 2: SMD Estimation Using SAC-SMA-SM and MODIS-LST as inputs**

363 Land surface temperature is the product of the soil temperature multiplied by the emissivity,

364 and the emissivity depends on the dielectric constant of the soil and soil moisture (Rodriguez-

365 Fernandez et al., 2015). Therefore the additional MODIS-LST information could potentially

366 improve the soil moisture estimation. The modelling process is the same as in Scheme 1. In

367 Table 4, it is clear to observe that by adding the MODIS-LST input, the Γ is improved to 0.060

368 and its corresponding gradient A is reduced significantly to less than half of the Scheme 1's.

369 Meanwhile the SE value is decreased remarkably as well showing the accuracy of the GT. The

370 M-test in Figure 4b shows the graph settles to an asymptote around 0.060 which is consistent

371 with the calculated Γ result. Training data size of 199 is chosen here because it gives the lowest

372   Γ value. For the LLR modelling, the best p_{max} value is found to be 2 from the trial and error

373   result in Figure 5. The LLR training and testing performances are presented in Figure 6b.

374   Although the problem of underestimation of extremely dry soil still exists (i.e., the points

375   concentrate at the right end of the training and testing plots), overall the model's prediction

376   ability during both phases are better than Scheme 1's (i.e., data points are closer to the 45° line).

377   The improvement can also be seen clearly in the time series plot in Figure 7b. For example, the

378   big disparities between the estimated and the targeted SMDs around DAY 300 and DAY 350

379   are reduced evidently.

3.3 Scheme 3: SMD Estimation Using SAC-SMA-SM and SMOS-T$_b$s as inputs

381   The multi-angle T$_b$s retrievals are the main data inputs for SMOS soil moisture calculation,

382   therefore their inclusion should also add a positive effect to the hydrological soil moisture

383   estimation. As aforementioned, an efficient feature selection of the SMOS incidence angles is

384   important for the best SMD calculation. In this study all the possible combinations from all

385   inputs variables are examined with the Γ result as the statistical indicator. This method is

386   capable of examining every combination (16383 embeddings in this case) of data inputs to

387   target the optimal combination that gives the smallest absolute Γ value. As discussed in Section

388   2.3.4, SAC-SMA-SM is a compulsory data input, so it is not included in the selecting process.

389   The best set of SMOS-T$_b$s to retrieve soil moisture state is composed of H polarization at the

390   incidence angles of 27.5^o-47.5^o, 57.5^o, and V polarization at the incidence angles of 27.5^o-42.5^o,

391   52.5^o, 57.5^o. This result demonstrates that using a combination of H and V T$_b$s gives a better

392   soil moisture estimation, which is logically sensible because different polarizations carry

393 distinct information of the Earth surface. However some incidence angles could held common

394 features which when putting together could result in a negative effect to the LLR modelling,

395 and are therefore not included. The detailed investigation of the possible common features is

396 out of the scope of this paper which is mainly due to the SMOS working mechanism.

397 As seen from Table 4, the inclusion of SMOS-T_bs significantly improves the Γ result by 54%,

398 while the gradient A is reduced greatly by 89% as compared with Scheme 1. The small A value

399 illustrates that the underlying system is more straightforward and easier to model than the

400 Scheme 1's. The M-test analysis in Figure 4c produces an asymptotic convergence from 120

401 training data size of Γ value around 0.033. It is interesting to see that the proportion of the

402 required training data is relatively larger than those in Scheme 1 and 2. The potential reason

403 could be explained by the larger amount of data inputs in this scheme. For LLR modelling, the

404 p_{max} that gives the smallest Γ is 7 (Figure 5). The SMD estimations during the training and the

405 testing are presented in Figure 8a. It can be seen that the SMD prediction ability of this scheme

406 is remarkably better than the previous ones, as most of the points lie on the bisector line albeit

407 there are still some under- and over- estimations. The reason SMOS outperforms MODIS in

408 SMD estimation could be due to the long wavelength microwave has, so it presents the top few

409 centimetres of the soil while MODIS LST (thermal infrared) only provides information at the

410 soil surface. The used LLR algorithm has been double checked to filter out the potential of

411 overfitting problem. The checking processes are performed by muddling the SMD target in the

412 testing datasets as well as altering the input file, and its efficiency stays the same. Hence it is

413 believed that the LLR model is very useful in calculating SMD from this scheme. Generally

414 the *NSE*, *r* and *RMSE* statistical indicators show a high agreement during both training and

415   testing phases. For the time series plot in Figure 7c, it is clear to see that most of the estimated

416   points lie closely to the benchmark line. The observed outliers could be partly due to the data

417   shortage in this scheme, so that not all the scenarios are covered in the datasets.

418   **3.4 Scheme 4: SMD Estimation Using SAC-SMA-SM, MODIS-LST, and SMOS-T_bs as**

419   **inputs**

420   In this scheme, all the three data sources are used to test if the modelling performance can be

421   further improved. Here the full embedding calculation is again carried out to explore the most

422   suitable incidence angles from the SMOS-T_bs. This is because the added MODIS-LST data

423   could carry identical (i.e., redundant) features with some of the SMOS-T_bs datasets. As a result

424   of the full embedding calculation, the best set of SMOS-T_bs is composed of H polarization at

425   the incidence angles of 37.5°-57.5°, and V polarization at the incidence angles of 37.5°-42.5°,

426   57.5°. As seen in Figure 4d, the total amount of data is significantly reduced due to the shortage

427   of simultanuously available days between the MODIS and the SMOS observations.

428   Interestingly the *M*-test graph vibrates more significantly than the other three schemes, which

429   could be due to the smaller data size and the larger amount of data inputs in this scheme. Here

430   the training data size is chosen as 62 with Γ obtained at around 0.030. The optimal p_{max} is

431   identified to be 5 (Figure 5). The LLR modelling results are shown in Figure 7d and Figure 8b.

432   It is obvious that this scheme further improves the accuracy of the SMD estimation, especially

433   with the high statistical performances achieved during both training and testing phases.

434   Comparatively this scheme is more stable for SMD estimation, albeit it requires more data

435   inputs and is only realisable when both the MODIS and the SMOS observations are available.

21

3.5 Produce an Unintermitted Soil Moisture Product

The data availability of the four schemes varies. As shown in Figure 9, Scheme 1 which has the poorest soil moisture state estimation gives the most data availability, while Scheme 4 which has the most accurate soil moisture state estimation owns the least data availability. In order to produce an unintermitted hydrological soil moisture product, the four schemes need to be combined together to complement each other. The combining method is by selecting the best available soil moisture estimation. For example if all the schemes have available data at the same time, the best scheme's soil moisture data is chosen (i.e., scheme 4 in this situation); whereas if just one scheme has data on that day, only that scheme's soil moisture data is used. The performances of the four schemes as well as the combined product are summarised in Table 5. Although the combined soil moisture state is obtained with lower statistical performances than Scheme 3's and 4's, it is still hydrologically very accurate especially when comparing with the SMOS's official soil moisture product (Table 5). The time series of the combined soil moisture state is plotted in Figure 10. It can be seen that the general trend of the produced soil moisture state follows the targeted data very well. However it tends to overestimate some of the wet events during the rainy season and significantly underestimate the dryer soil condition in September 2011. Those poor estimations are mostly from the Scheme 1 and 2 where Schemes 3 and 4 are not available. Since more and more microwave satellite observations are becoming obtainable, those new data sources could add extra benefits into the proposed model, and the accuracy of the soil moisture product is expected to be further enhanced.

22

457 **4. Discussion**

458 - *What is a soil moisture state variable?*

459 This study uses the XAJ's SMD simulation as a target because it is hydrological model directly

460 produced. However it is argued that models with different parameters values can generate

461 equally good flow results named as the equifinality effect, because they are all calibrated based

462 on the observed flow. For this reason, their soil moisture state variables can be distinct among

463 each other.

464 In order to investigate this effect in more details, the XAJ model is manipulated by increasing

465 one of its parameters *WUM* by 30 %. By doing so, the XAJ's flow simulation remains as

466 effective as its original form (the same *NSE* values), but its soil moisture state changes

467 significantly from its original values. For a better visualisation, an enlarged plot of the SMD

468 simulations between Day 222 and Day 344 is presented. As seen from Figure 11a although the

469 soil moisture state variables from two equally good calibrations have a wide range of value

470 differences ($NSE = 0.34$), they both follow the same pattern: when it rains they become wet by

471 the similar amount; when there is a dry period they all move into a dryer state in a similar rate

472 to the actual evapotranspiration. Therefore they appear as in parallel movements and the latter

473 plot (Figure 11b) shows a very strong linear correlation ($r = 1.0$) between them. It is important

474 to note that the selection of the dry period (i.e., high SMD values) is because it is the most

475 critical period of time for the need of accurate soil moisture values for hydrological modelling.

476 This is because during the real-time flood forecasting, after a long period of dryness, the

477 accumulation of error in the hydrological models can become larger and larger with time. With

478 accurate soil moisture information, the error could be corrected.

479 Although the absolute values of the models' soil moisture state variables are not quite

480 meaningful and comparable, their variations are the true reflection of the soil moisture

481 fluctuations in the real-world. This clarification is a very important concept, because there has

482 been a wide spread of misunderstanding about the hydrological model's soil moisture state and

483 its connection with the real-world soil moisture.

484 - *Soil moisture state normalisation*

485 One deficiency of this study is that the generated soil moisture state is based on a hydrological

486 model's SMD simulation, so it is model parameter dependent. It is desirable to produce a soil

487 moisture indicator which is independent from model parameters and dimensionless with

488 variables between 0 and 1. Normalised Hydrological Soil Moisture State (NHSMS) indicators

489 are produced as presented in Figure 12 (corresponding to the SMD simulations shown in Figure

490 11). The normalisation method is by adopting the following equation:

491 $$NHSMS = \frac{SMD - \min(SMD)}{\max(SMD) - \min(SMD)} \qquad (11)$$

492 Such an approach is very effective as demonstrated by the almost identical SMD curves

493 between the two XAJ simulations. In the future it is planned to use the same process on other

494 hydrological models to test if the normalised soil moisture indicators are not only model

495 parameter independent but also model structure independent. Since all hydrological models are

496 driven by the same physics laws on the conservation of mass, their normalised soil moisture

497   indicators should respond in a similar way (soil becomes wetter when it rains and drier when

498   there is no rain). If this is true a new soil moisture product based on NHSMS could be generated

499   as a routine product by the operational organisations such as NASA and ESA. Such a soil

500   moisture product will also be very useful to the meteorological and hydro-meteorological fields

501   in their land surface modelling because the current land surface models suffer from poor

502   performance in their runoff estimations. As aforementioned, all current soil moisture products

503   such as those from ESA and NASA are not optimised for different application fields. Our study

504   gives an example of simulating the soil moisture data targeted to serve the hydrological

505   community. It is possible other products serving farmers in agriculture, ecologists in the

506   environment, and geotechnical engineers in construction could be produced using the proposed

507   method.

508   - *Application of the produced soil moisture data*

509   Another area needs further work is the hydrological application of the produced data. Generally

510   effective hydrological application of soil moisture data needs three pre-conditions: 1) a good

511   soil moisture data relevant to hydrology; 2) a hydrological model compatible with such data;

512   3) an effective data assimilation scheme. This paper tackles the first point, and the other two

513   points would need further research because there are significant knowledge gaps in them. If all

514   the three points are solved, such a data has a huge potential in operational hydrological

515   modelling. For example, initialisation of the model could be shortened which reduces the need

516   for model warm up. This is important during real-time flood forecasting when there is not

517   enough data to warm up the model for an imminent flood event. Such a warm-up period could

518   be very long, as demonstrated by the study in (Ceola et al., 2015). In addition the XAJ SMD

519   data used here is based on the calibration of the observed rainfall and flow, so that the targeted

520   SMD is interpolated between observations and there is a minimum time-drift. In the real-time

521   flood forecasting the errors in precipitation and evapotranspiration could accumulate which

522   cause time-drift problems. Therefore a soil moisture product such as the one produced in this

523   study (i.e., based on minimal time-drift SMD) could help avoiding such a problem. The

524   proposed soil moisture data is also valuable for the validation of land surface models, especially

525   useful for their runoff simulations. Due to the limit of time and resources this study has not

526   tackled all the issues, but has laid a good foundation for their future researches.

527   **5. Conclusions**

528   A hydrological soil moisture product is produced for the Pontiac catchment using the GT and

529   the LLR modelling techniques based on four data-input schemes. Three data sources are

530   considered including the soil moisture product from the SAC-SMA model, the land surface

531   temperature retrieved by the MODIS satellite, and the multi-angle brightness temperatures

532   acquired from the SMOS satellite. The four data-input schemes are built from the four

533   combinations of the data sources. The generated soil moisture product (unintermitted with no

534   missing data) for a period of two years (2010-2011) is compared with the XAJ hydrological

535   model's SMD simulation to test its hydrological accuracy. It is concluded that the GT and the

536   LLR modelling techniques together with the chosen data inputs can be used with high

537   confidence to estimate an unintermitted hydrological soil moisture product, and the proposed

538   method could be easily applied to other catchments and fields.

539 In this study it has been found that different data sources have their own unique information

540 contents, so that they can complement each other using data fusion technique. Their synergy

541 can be best achieved to produce an enhanced soil moisture product. In data fusion an important

542 principle is MRmr (Maximum Relevance minimum redundancy). The soil moisture state in

543 this study is generated from a large number of data inputs, and their selection is carried out by

544 the GT which is one of the methods in MRmr. This is the first time that the GT is used in a data

545 fusion of satellite multiple T_bs scans, land surface temperature and external soil moisture

546 information for producing a hydrological soil moisture product. Future studies should explore

547 other MRmr methods in addition to GT, to compare if they are more effective input selection

548 methods. As to the data fusion regression model, LLR is chosen in this study because it is easily

549 applied and very effective. However it is possible there may exist other better models. We

550 encourage the community to apply the proposed methodology using other regression models.

551 **Acknowledgments**

27

Level-3 land surface temperature can be obtained from the LP DAAC website at

https://lpdaac.usgs.gov/dataset_discovery/modis/modis_products_table/mod11c1.

References

Al-Bitar, Ahmad, Leroux, Delphine, Kerr, Yann H, Merlin, Olivier, Richaume, Philippe, Sahoo, Alok, Wood, Eric F, 2012. Evaluation of SMOS soil moisture products over continental US using the SCAN/SNOTEL network. Geoscience and Remote Sensing, IEEE Transactions on, 50(5): 1572-1586.

Aubert, David, Loumagne, Cecile, Oudin, Ludovic, 2003. Sequential assimilation of soil moisture and streamflow data in a conceptual rainfall–runoff model. Journal of Hydrology, 280(1): 145-161.

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

Berg, Aaron A, Mulroy, Kathleen A, 2006. Streamflow predictability in the Saskatchewan/Nelson River basin given macroscale estimates of the initial soil moisture status. Hydrological sciences journal, 51(4): 642-654.

Berthet, L, Andréassian, V, Perrin, C, Javelle, P, 2009. How crucial is it to account for the antecedent moisture conditions in flood forecasting? Comparison of event-based and continuous approaches on 178 catchments. Hydrology and Earth System Sciences Discussions(13): p. 819-p. 831.

Beven, Keith J, 2012. Rainfall-runoff modelling: the primer. John Wiley & Sons.

Brocca, L, Melone, F, Moramarco, T, Wagner, W, Naeimi, V, Bartalis, Z, Hasenauer, S, 2010. Improving runoff prediction through the assimilation of the ASCAT soil moisture product. Hydrology and Earth System Sciences Discussions, 7(4): 4113-4144.

583 Calder, IR, Harding, RJ, Rosier, PTW, 1983. An objective assessment of soil-moisture deficit
584   models. Journal of Hydrology, 60(1): 329-355.

585 Carlson, Toby, 2007. An overview of the" triangle method" for estimating surface
586   evapotranspiration and soil moisture from satellite imagery. Sensors, 7(8): 1612-1629.

587 Ceola, Serena, Arheimer, Berit, Baratti, E, Blöschl, G, Capell, Rene, Castellarin, Attilio, Freer,
588   Jim, Han, Dawei, Hrachowitz, Markus, Hundecha, Yeshewatesfa, 2015. Virtual
589   laboratories: new opportunities for collaborative water science. Hydrology and Earth
590   System Sciences, 19(4): 2101-2117.

591 Chen, Fan, Crow, Wade T, Starks, Patrick J, Moriasi, Daniel N, 2011. Improving hydrologic
592   predictions of a catchment model via assimilation of surface soil moisture. Advances
593   in Water Resources, 34(4): 526-536.

594 Chen, Jieyun, Adams, Barry J, 2006. Integration of artificial neural networks with conceptual
595   models in rainfall-runoff modeling. Journal of Hydrology, 318(1): 232-249.

596 Chen, Xi, Yang, Tao, Wang, Xiaoyan, Xu, Chong-Yu, Yu, Zhongbo, 2013. Uncertainty
597   Intercomparison of Different Hydrological Models in Simulating Extreme Flows.
598   Water resources management, 27(5): 1393-1409.

599 Dumedah, G, Coulibaly, P, 2013. Evolutionary assimilation of streamflow in distributed
600   hydrologic modeling using in-situ soil moisture data. Advances in Water Resources, 53:
601   231-241.

602 Durrant, PJ, 2001. winGammaTM: A non-linear data analysis and modelling tool for the
603   investigation of non-linear and chaotic systems with applied techniques for a flood
604   prediction system. PhD Thesis, Cardiff University.

605 Eltahir, Elfatih AB, 1998. A soil moisture-rainfall feedback mechanism 1. Theory and
606   observations. Water Resources Research, 34(4): 765-776.

607   Entekhabi, Dara, Rodriguez-Iturbe, Ignacio, 1994. Analytical framework for the
608   characterization of the space-time variability of soil moisture. Advances in water
609   resources, 17(1): 35-45.

610   ESA Earth Online, 2016. SMOS soil moisture product in NRT based on neural network is now
611   available, https://earth.esa.int/web/guest/missions/esa-operational-eo-
612   missions/smos/news/-/article/smos-soil-moisture-product-in-nrt-based-on-neural-
613   network-is-now-available. Accessed on 13/10/2016.

614   Evans, Dafydd, Jones, Antonia J, 2002. A proof of the Gamma test, Proceedings of the Royal
615   Society of London A: Mathematical, Physical and Engineering Sciences. The Royal
616   Society, pp. 2759-2799.

617   Gan, Thian Yew, Dlamini, Enoch M, Biftu, Getu Fana, 1997. Effects of model complexity and
618   structure, data quality, and objective functions on hydrologic modeling. Journal of
619   Hydrology, 192(1): 81-103.

620   Goward, Samuel N, Xue, Yongkang, Czajkowski, Kevin P, 2002. Evaluating land surface
621   moisture conditions from the remotely sensed temperature/vegetation index
622   measurements: An exploration with the simplified simple biosphere model. Remote
623   sensing of environment, 79(2): 225-242.

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

627   Jaafar, WZ Wan, Han, D, 2011. Variable selection using the gamma test forward and backward
628   selections. Journal of Hydrologic Engineering, 17(1): 182-190.

629   Jacquette, Elsa, Al Bitar, Ahmad, Mialon, Arnaud, Kerr, Yann, Quesney, Arnaud, Cabot,
630   François, Richaume, Philippe, 2010. SMOS CATDS level 3 global products over land,
631   Remote Sensing for Agriculture, Ecosystems, and Hydrology XII. International Society
632   for Optics and Photonics, Toulouse, France DOI:10.1117/12.865093

633   Javelle, Pierre, Fouchier, Catherine, Arnaud, Patrick, Lavabre, Jacques, 2010. Flash flood
634   warning at ungauged locations using radar rainfall and antecedent soil moisture
635   estimations. Journal of hydrology, 394(1): 267-274.

636   Kerr, Yann H, Waldteufel, Philippe, Richaume, Philippe, Wigneron, J-P, Ferrazzoli, Paolo,
637   Mahmoodi, Ali, Al Bitar, Ahmad, Cabot, François, Gruhier, Claire, Juglea, Silvia
638   Enache, 2012. The SMOS soil moisture retrieval algorithm. Geoscience and Remote
639   Sensing, IEEE Transactions on, 50(5): 1384-1403.

640   Kerr, Yann H, Waldteufel, Philippe, Wigneron, J-P, Delwart, Steven, Cabot, François, Boutin,
641   Jacqueline, Escorihuela, M-J, Font, Jordi, Reul, Nicolas, Gruhier, Claire, 2010a. The
642   SMOS mission: New tool for monitoring key elements ofthe global water cycle.
643   Proceedings of the IEEE, 98(5): 666-687.

644   Kerr, Yann H, Waldteufel, Philippe, Wigneron, J-P, Martinuzzi, J, Font, Jordi, Berger, Michael,
645   2001. Soil moisture retrieval from space: The Soil Moisture and Ocean Salinity (SMOS)
646   mission. Geoscience and Remote Sensing, IEEE Transactions on, 39(8): 1729-1735.

647   Kerr, Yann H, Waldteufel, Philippe, Wigneron, Jean-Pierre, Delwart, Steven, Cabot, François
648   Ois, Boutin, Jacqueline, Escorihuela, Maria-José, Font, Jordi, Reul, Nicolas, Gruhier,
649   Claire, 2010b. The SMOS mission: New tool for monitoring key elements ofthe global
650   water cycle. Proceedings of the IEEE, 98(5): 666-687.

651   Koncar, N, 1997. Optimisation methodologies for direct inverse neurocontrol. PhD thesis
652   Thesis, University of London, Imperial College of Science, Technology and Medicine,
653   London, SW7 2BZ.

654   Liang, Xetal, Lettenmaier, Dennis P, Wood, Eric F, Burges, Stephen J, 1994. A simple
655   hydrologically based model of land surface water and energy fluxes for general
656   circulation models. Journal of Geophysical Research: Atmospheres, 99(D7): 14,415-
657   14,428.

658 Liu, Xianming, Zhao, Debin, Xiong, Ruiqin, Ma, Siwei, Gao, Wen, Sun, Huifang, 2011. Image
659   interpolation via regularized local linear regression. Image Processing, IEEE
660   Transactions on, 20(12): 3455-3469.

661 Mallick, Kaniska, Bhattacharya, Bimal K, Patel, NK, 2009. Estimating volumetric surface
662   moisture content for cropped soils using a soil wetness index based on surface
663   temperature and NDVI. Agricultural and Forest Meteorology, 149(8): 1327-1342.

664 Matgen, P, Heitz, S, Hasenauer, S, Hissler, C, Brocca, L, Hoffmann, L, Wagner, W, Savenije,
665   HHG, 2012a. On the potential of MetOp ASCAT-derived soil wetness indices as a new
666   aperture for hydrological monitoring and prediction: a field evaluation over
667   Luxembourg. Hydrological Processes, 26(15): 2346-2359.

668 Matgen, Patrick, Fenicia, Fabrizio, Heitz, Sonia, Plaza, Douglas, de Keyser, Robain, Pauwels,
669   Valentijn RN, Wagner, Wolfgang, Savenije, Hubert, 2012b. Can ASCAT-derived soil
670   wetness indices reduce predictive uncertainty in well-gauged areas? A comparison with
671   in situ observed soil moisture in an assimilation application. Advances in Water
672   Resources, 44: 49-65.

673 Mitchell, Kenneth E, Lohmann, Dag, Houser, Paul R, Wood, Eric F, Schaake, John C, Robock,
674   Alan, Cosgrove, Brian A, Sheffield, Justin, Duan, Qingyun, Luo, Lifeng, 2004. The
675   multi-institution North American Land Data Assimilation System (NLDAS): Utilizing
676   multiple GCIP products and partners in a continental distributed hydrological modeling
677   system. Journal of Geophysical Research: Atmospheres (1984–2012), 109(D7).
678   DOI:10.1029/2003JD003823

679 MODAPS Services, 2015. Terra Product Descriptions: MOD11_L2,
680   http://modaps.nascom.nasa.gov/services/about/products/c6-nrt/MOD11_L2.html.
681   Accessed on 13/10/2016.

682 Moore, RJ, 2007. The PDM rainfall-runoff model. Hydrology and Earth System Sciences
683   Discussions, 11(1): 483-499.

32

684   Nash, JEa, Sutcliffe, JV, 1970. River flow forecasting through conceptual models part I—A
685   discussion of principles. Journal of Hydrology, 10(3): 282-290.

686   Njoku, Eni G, Entekhabi, Dara, 1996. Passive microwave remote sensing of soil moisture.
687   Journal of hydrology, 184(1): 101-129.

688   Njoku, Eni G, Kong, Jin-Au, 1977. Theory for passive microwave remote sensing of near-
689   surface soil moisture. Journal of Geophysical Research, 82(20): 3108-3118.

690   Noori, R, Karbassi, AR, Moghaddamnia, A, Han, D, Zokaei-Ashtiani, MH, Farokhnia, A,
691   Gousheh, M Ghafari, 2011. Assessment of input variables determination on the SVM
692   model performance using PCA, Gamma test, and forward selection techniques for
693   monthly stream flow prediction. Journal of Hydrology, 401(3): 177-189.

694   Norbiato, Daniele, Borga, Marco, Degli Esposti, Silvia, Gaume, Eric, Anquetin, Sandrine,
695   2008. Flash flood warning based on rainfall thresholds and soil moisture conditions:
696   An assessment for gauged and ungauged basins. Journal of Hydrology, 362(3): 274-
697   290.

698   Peel, Murray C, Finlayson, Brian L, McMahon, Thomas A, 2007. Updated world map of the
699   Köppen-Geiger climate classification. Hydrology and Earth System Sciences
700   Discussions, 4(2): 439-473.

701   Peng, G., Leslie, L.M., Shao, Y., 2002. Environmental Modelling and Prediction. Springer.
702   480 pp.

703   Penrose, Roger, 1955. A generalized inverse for matrices, Mathematical proceedings of the
704   Cambridge philosophical society. Cambridge Univ Press, pp. 406-413.

705   Penrose, Roger, 1956. On best approximate solutions of linear matrix equations, Mathematical
706   Proceedings of the Cambridge Philosophical Society. Cambridge Univ Press, pp. 17-
707   19.

708   Pi, Hong, Peterson, Carsten, 1994. Finding the embedding dimension and variable
709   dependencies in time series. Neural Computation, 6(3): 509-520.

710 Pierdicca, Nazzareno, Pulvirenti, Luca, Bignami, Christian, Ticconi, Francesca, 2013.
711   Monitoring soil moisture in an agricultural test site using SAR data: design and test of
712   a pre-operational procedure. Selected Topics in Applied Earth Observations and
713   Remote Sensing, IEEE Journal of, 6(3): 1199-1210.

714 Pinson, Pierre, Nielsen, Henrik Aa, Madsen, Henrik, Nielsen, Torben S, 2008. Local linear
715   regression with adaptive orthogonal fitting for the wind power application. Statistics
716   and Computing, 18(1): 59-71.

717 Prakash, Rishi, Singh, Dharmendra, Pathak, Nagendra P, 2012. A fusion approach to retrieve
718   soil moisture with SAR and optical data. Selected Topics in Applied Earth Observations
719   and Remote Sensing, IEEE Journal of, 5(1): 196-206.

720 Price, John C, 1980. The potential of remotely sensed thermal infrared data to infer surface soil
721   moisture and evaporation. Water Resources Research, 16(4): 787-795.

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

724 Rodell, Matthew, 2016. NLDAS Concept/Goals, NLDAS Concept/Goals,
725   http://ldas.gsfc.nasa.gov/nldas/NLDASgoals.php. Accessed on 13/10/2016.

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

730 Rodriguez-Fernandez, Nemesio J, Aires, Filipe, Richaume, Philippe, Kerr, Yann H, Prigent,
731   Catherine, Kolassa, Jana, Cabot, Francois, Jimenez, Carlos, Mahmoodi, Ali, Drusch,
732   Matthias, 2015. Soil moisture retrieval using neural networks: application to SMOS.
733   IEEE Transactions on Geoscience and Remote Sensing, 53(11): 5991 - 6007.

734 Romano, Nunzio, 2014. Soil moisture at local scale: Measurements and simulations. Journal
735   of Hydrology, 516: 6-20.

736 Rushton, KR, Eilers, VHM, Carter, RC, 2006. Improved soil moisture balance methodology
737   for recharge estimation. Journal of Hydrology, 318(1): 379-399.

738 Shi, Peng, Chen, Chao, Srinivasan, Ragahavan, Zhang, Xuesong, Cai, Tao, Fang, Xiuqin, Qu,
739   Simin, Chen, Xi, Li, Qiongfang, 2011. Evaluating the SWAT model for hydrological
740   modeling in the Xixian watershed and a comparison with the XAJ model. Water
741   resources management, 25(10): 2595-2612.

742 Srivastava, PK, Han, D, Ramirez, MR, Islam, T, 2013. Machine Learning Techniques for
743   Downscaling SMOS Satellite Soil Moisture Using MODIS Land Surface Temperature
744   for Hydrological Application. Water Resources Management, 27(8): 3127–3144.

745 Stefánsson, Adoalbjörn, Končar, N, Jones, Antonia J, 1997. A note on the gamma test. Neural
746   Computing & Applications, 5(3): 131-133.

747 Sun, Hongyu, Liu, Henry, Xiao, Heng, He, Rachel, Ran, Bin, 2003. Use of local linear
748   regression model for short-term traffic forecasting. Transportation Research Record:
749   Journal of the Transportation Research Board(1836): 143-150.

750 Todini, E, 1996. The ARNO rainfall—runoff model. Journal of Hydrology, 175(1): 339-382.

751 Tsui, Alban PM, Jones, Antonia J, De Oliveira, A Guedes, 2002. The construction of smooth
752   models using irregular embeddings determined by a gamma test analysis. Neural
753   Computing & Applications, 10(4): 318-329.

754 Wagner, W, Dorigo, Wouter, de Jeu, Richard, Fernandez, Diego, Benveniste, Jerome, Haas,
755   Eva, Ertl, Martin, 2012. Fusion of active and passive microwave observations to create
756   an essential climate variable data record on soil moisture, Proceedings of the XXII
757   International Society for Photogrammetry and Remote Sensing (ISPRS) Congress,
758   Melbourne, Australia.

759 Wan, Zhengming, 2008. New refinements and validation of the MODIS land-surface
760   temperature/emissivity products. Remote Sensing of Environment, 112(1): 59-74.

761 Wan, Zhengming, Zhang, Yulin, Zhang, Qincheng, Li, Zhao-liang, 2002. Validation of the

762   land-surface temperature products retrieved from Terra Moderate Resolution Imaging

763   Spectroradiometer data. Remote sensing of Environment, 83(1): 163-180.

764 Wanders, Niko, Bierkens, Marc FP, de Jong, Steven M, de Roo, Ad, Karssenberg, Derek, 2014.

765   The benefits of using remotely sensed soil moisture in parameter identification of large-

766   scale hydrological models. Water Resources Research, 50(8): 6874-6891.

767 Wang, QJ, 1991. The genetic algorithm and its application to calibrating conceptual rainfall-

768   runoff models. Water resources research, 27(9): 2467-2471.

769 Webb, Robert W, Rosenzweig, Cynthia E, Levine, Elissa R, 2000. Global Soil Texture and

770   Derived Water-Holding Capacities (Webb et al.). Data set. Available on-line

771   [http://www. daac. ornl. gov] from Oak Ridge National Laboratory Distributed Active

772   Archive Center, Oak Ridge, Tennessee, USA.

773 Xia, Youlong, Mitchell, Kenneth, Ek, Michael, Sheffield, Justin, Cosgrove, Brian, Wood, Eric,

774   Luo, Lifeng, Alonge, Charles, Wei, Helin, Meng, Jesse, 2012. Continental-scale water

775   and energy flux analysis and validation for the North American Land Data Assimilation

776   System project phase 2 (NLDAS-2): 1. Intercomparison and application of model

777   products. Journal of Geophysical Research: Atmospheres (1984–2012), 117(D3).

778 Xia, Youlong, Sheffield, Justin, Ek, Michael B, Dong, Jiarui, Chaney, Nathaniel, Wei, Helin,

779   Meng, Jesse, Wood, Eric F, 2014. Evaluation of multi-model simulated soil moisture

780   in NLDAS-2. Journal of Hydrology, 512: 107-125.

781 Zhao, RenJun, 1992. The Xinanjiang model applied in China. Journal of Hydrology, 135(1):

782   371-381.

783 Zhao, RenJun, Liu, XR, 1995. The Xinanjiang model. In: Singh, V.P. (Ed.), Computer models

784   of watershed hydrology., pp. 215-232.

785 Zhuo, Lu, Dai, Qiang, Han, Dawei, 2015a. Evaluation of SMOS soil moisture retrievals over

786   the central United States for hydro-meteorological application. Physics and Chemistry

787   of the Earth, Parts A/B/C. DOI:10.1016/j.pce.2015.06.002

788   Zhuo, Lu, Dai, Qiang, Islam, Tanvir, Han, Dawei, 2016a. Error distribution modelling of

789   satellite soil moisture measurements for hydrological applications. Hydrological

790   Processes, 30(13): 2223-2236.

791   Zhuo, Lu, Han, Dawei, 2016a. Could operational hydrological models be made compatible

792   with satellite soil moisture observations? Hydrological Processes, 30(10): 1637-1648.

793   Zhuo, Lu, Han, Dawei, 2016b. Misrepresentation and amendment of soil moisture in

794   conceptual hydrological modelling. Journal of Hydrology, 535: 637-651.

795   Zhuo, Lu, Han, Dawei, Dai, Qiang, 2016b. Soil moisture deficit estimation using satellite

796   multi-angle brightness temperature. Journal of Hydrology, 539: 392-405.

797   Zhuo, Lu, Han, Dawei, Dai, Qiang, Islam, Tanvir, Srivastava, Prashant K, 2015b. Appraisal of

798   NLDAS-2 Multi-Model Simulated Soil Moistures for Hydrological Modelling. Water

799   Resources Management, 29(10): 3503-3517.

800

Table 1. The XAJ model parameters used in the Pontiac catchment.

Symbol	Model parameters	Unit	Range
K	Ratio of evapotranspiration	[-]	0.10-1.20
WUM	The areal mean field capacity of the upper layer	mm	30-50
WLM	The areal mean field capacity of the lower layer	mm	20-150
WDM	The areal mean field capacity of the deep layer	mm	30-400
IMP	Percentage of impervious and saturated areas in the catchment	%	0.00-0.10
B	Exponential parameter with a single parabolic curve, which represents the non-uniformity of the spatial distribution of the soil moisture storage capacity over the catchment	[-]	0.10-0.90
C	Coefficient of the deep layer that depends on the proportion of the catchment area covered by vegetation with deep roots	[-]	0.10-0.70
SM	Areal mean free water capacity, which represents the maximum possible deficit of free water storage	mm	10-50
KG	Outflow coefficient of the free water storage to groundwater relationships	[-]	0.10-0.70
KSS	Outflow coefficient of the free water storage to interflow relationships	[-]	0.10-0.70
EX	Exponent of the free water capacity curve	[-]	1.10-2.00
KKG	Recession constant of the groundwater storage	[-]	0.01-0.99
$KKSS$	Recession constant of the lower interflow storage	[-]	0.01-0.99
CS	Recession constant in the lag and route method for routing through the channel system with each sub-catchment	[-]	0.10-0.70
L	Lag in time	[-]	0.00-6.00
V	Parameter of the Muskingum method	m/s	0.40-1.20
dX	Parameter of the Muskingum method	[-]	0.00-0.40

Table 2. General data-input properties relevant for this study.

	SMOS-T_bs	MODIS-LST	SAC-SMA-SM
Product	brightness temperature	land surface temperature	soil moisture
Unit	Kelvin (K)	Kelvin (K)	m^3/m^3
Near-Real-Time (NRT)	Yes	Yes	Yes
Spatial resolution (km)	35-50	5.6	14
Data time-step	~ every three days	~ daily	Daily
Data availability for the studying period (days)	217	458	730

Table 3. Four data-input schemes: scheme 1: SAC-SMA-SM; scheme 2: SAC-SMA-SM and MODIS-LST; scheme 3: SAC-SMA-SM and SMOS-T_bs; scheme 4: SAC-SMA-SM, MODIS-LST, and SMOS-T_bs.

	SAC-SMA-SM	MODIS-LST	SMOS-T_bs
Scheme 1	x		
Scheme 2	x	x	
Scheme 3	x		x
Scheme 4	x	x	x

Table 4. Model statistical performances and modelling information, where Γ is the calculated gamma statistic which is the minimum *MSE* that can be achieved from a modelling method; *A* is the Gamma gradient; *SE* is the Standard error; p_{max} is the nearest points for LLR modelling; *M* is the training data size; and SMOS IA is the chosen incidence angles of SMOS-T_bs.

	Γ	*A*	*SE*	p_{max}	*M*	SMOS IA
Scheme 1	0.072	1.353	0.004	4	292	-
Scheme 2	0.060	0.568	0.002	2	199	-
Scheme 3	0.033	0.152	0.004	7	120	H: 27.5º-47.5º, 57.5º
						V: 27.5º-42.5º, 52.5º, 57.5º
Scheme 4	0.029	0.119	0.006	5	62	H: 37.5º-57.5º
						V: 37.5º-42.5º, 57.5º

Table 5. Summary of SMD estimation performances. It is noted that *RMSE* is in the unit of metre.

	Training			Testing		
	NSE	*r*	*RMSE*	*NSE*	*r*	*RMSE*
Scheme 1	0.752	0.870	0.011	0.688	0.830	0.014
Scheme 2	0.767	0.877	0.011	0.747	0.865	0.012
Scheme 3	0.928	0.965	0.006	0.876	0.940	0.008
Scheme 4	0.912	0.957	0.007	0.912	0.960	0.007
Combined	-	-	-	0.790	0.889	0.011
SMOS-SM	-	-	-	0.420	0.650	0.017
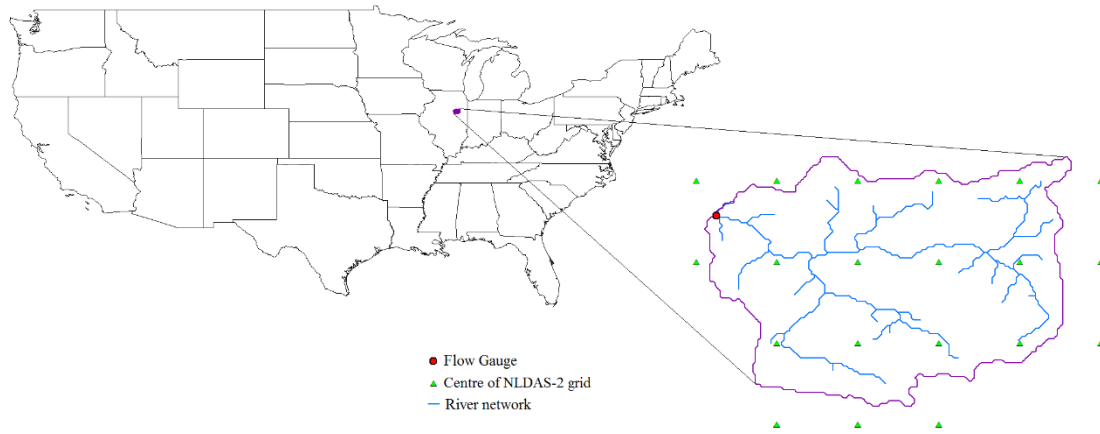
Figure. 1. The location and river network of the Pontiac catchment in the U.S., with the flow gauge and NLDAS-2 central grid points (Zhuo et al., 2015a).
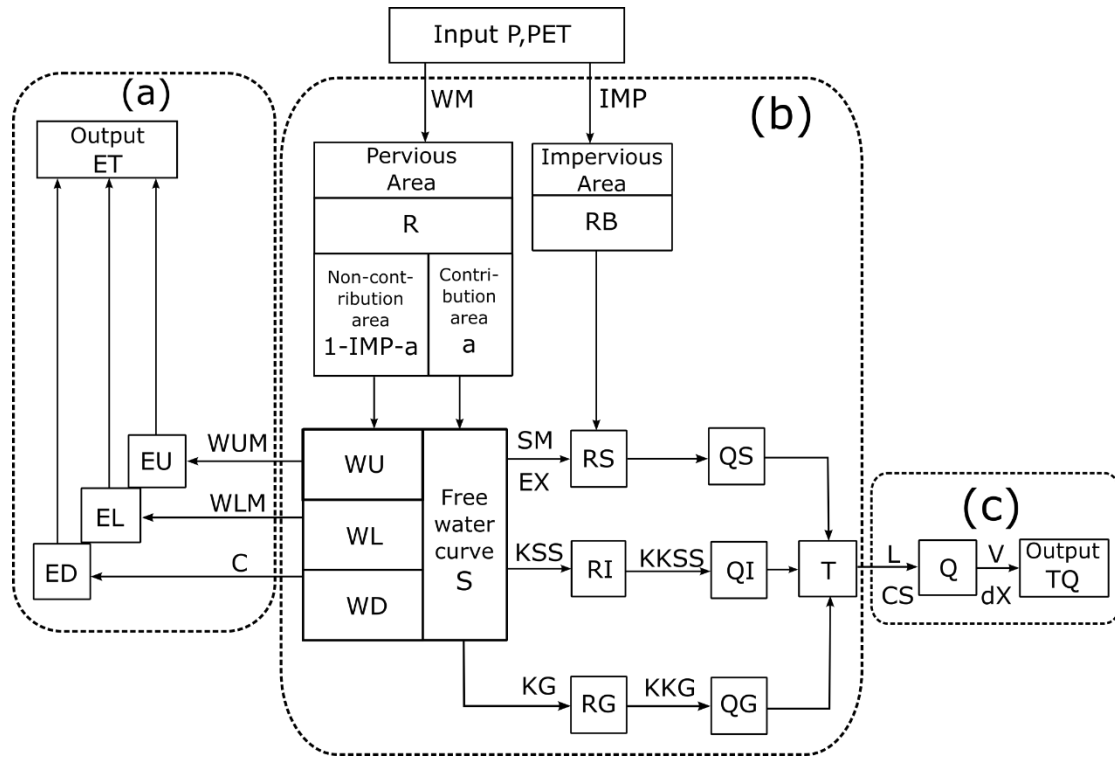
Figure. 2. Adopted flowchart of the XAJ model *(Zhao, 1992)*. The model consists of an evapotranspiration component (a), a runoff generating component (b), and a runoff routing component (c). *P*, *PET*, and *ET* are the precipitation, potential evapotranspiration, and the simulated actual evapotranspiration respectively; *WU*, *WL* and *WD* represent the upper, lower, and deep soil layers' areal mean tension water storage respectively; *WM* is the areal mean field capacity; *EU*, *EL,* and *ED* stand for the upper, lower, and deep soil layers' evapotranspiration output respectively; *S* is the areal mean free water storage; *a* is the portion of the sub-catchment producing runoff; *IMP* is the factor of impervious area in a catchment; *RB* is the direct runoff produced from the small portion of impervious area; *R* is the total runoff generated from the model with surface runoff (*RS*), interflow (*RI*), and groundwater runoff (*RG*) components respectively. These three runoff components are then transferred into *QS*, *QI,* and *QG* and combined as the total sub-catchment inflow (*T*) to the channel network. The flow outputs *Q* from each sub-catchment are then routed to the catchment outlet to produce the final flow result (*TQ*). The rest of the symbols are explained in Table 1.
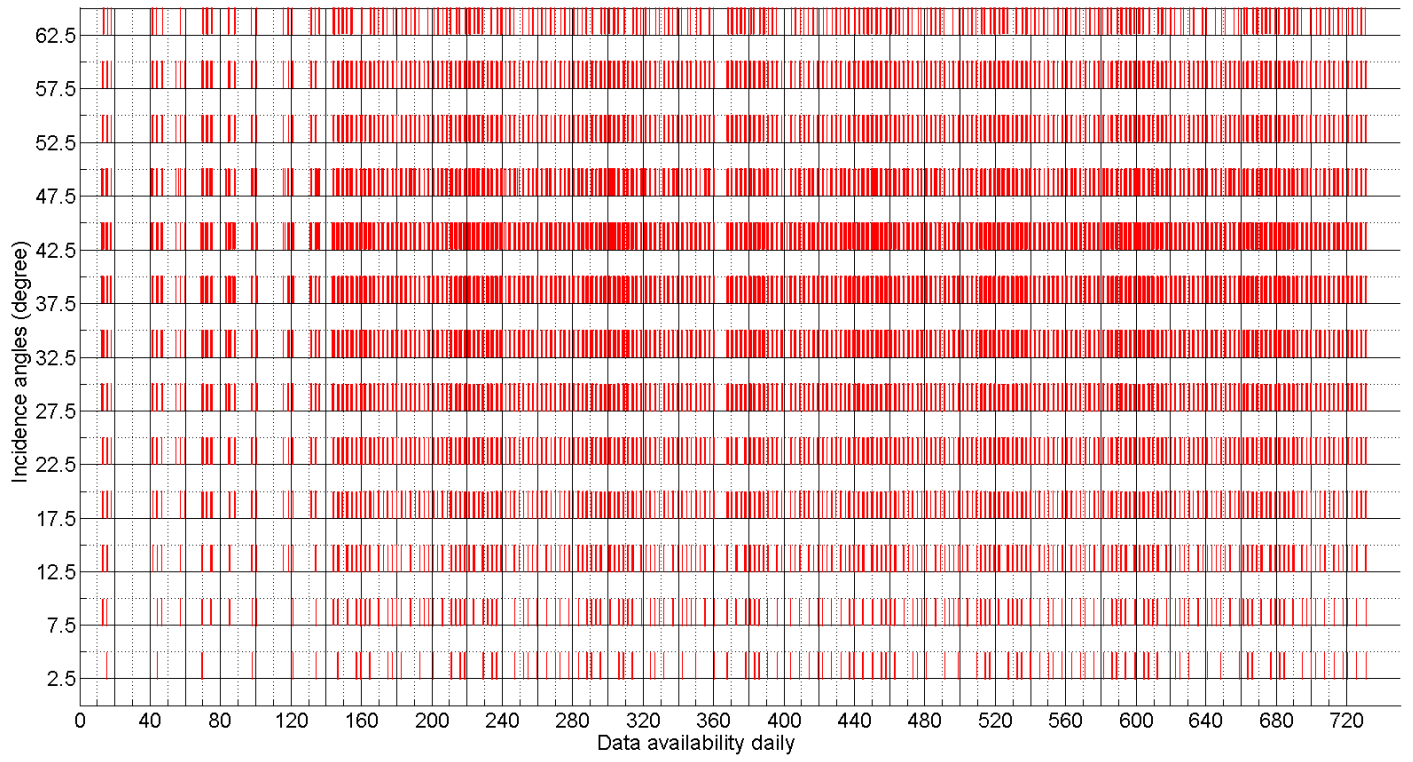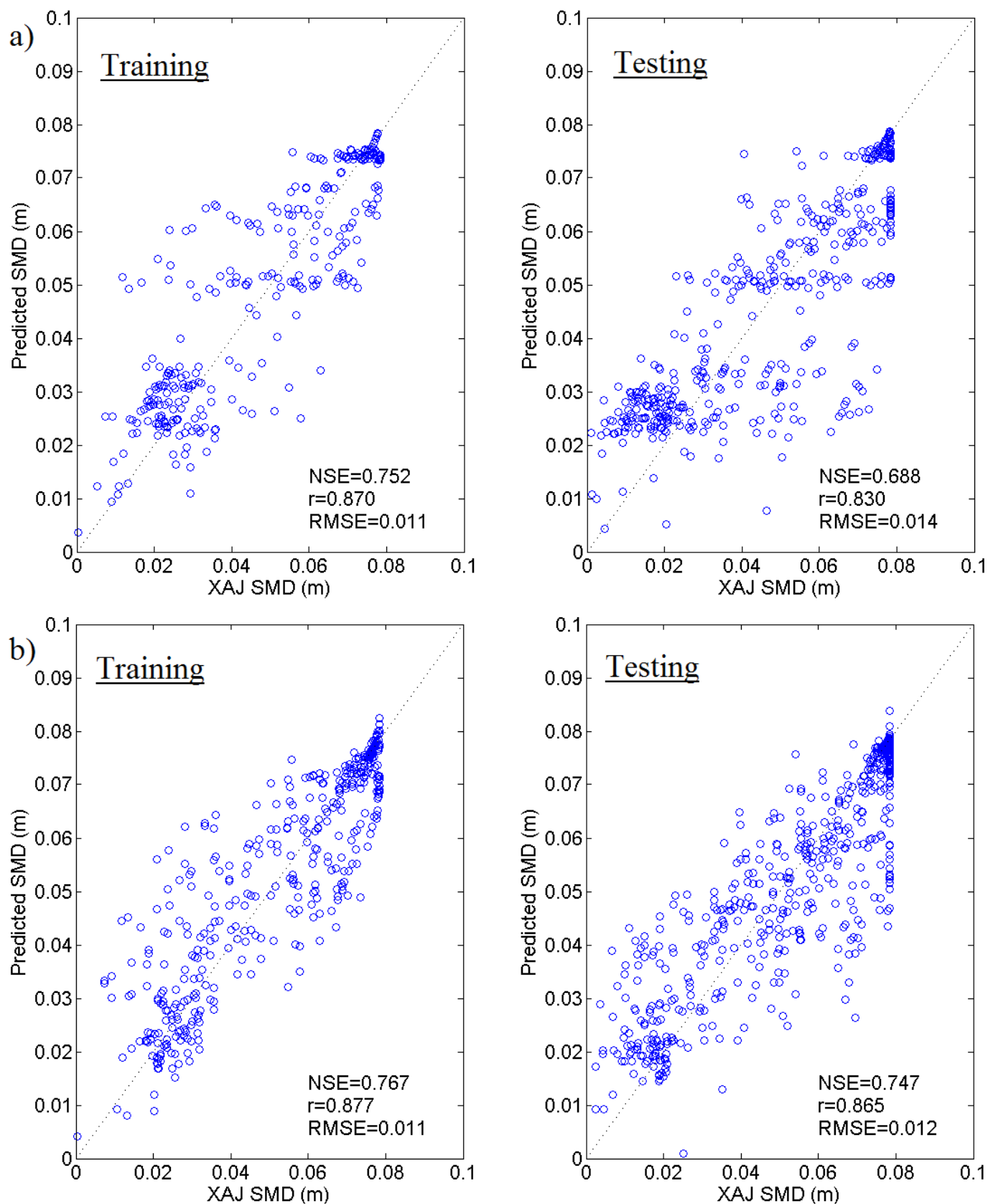
Figure 3. SMOS-T_bs data availabilities. It is noted that the available dates for the horizontal and the vertical polarizations are the same, so only one is shown here.
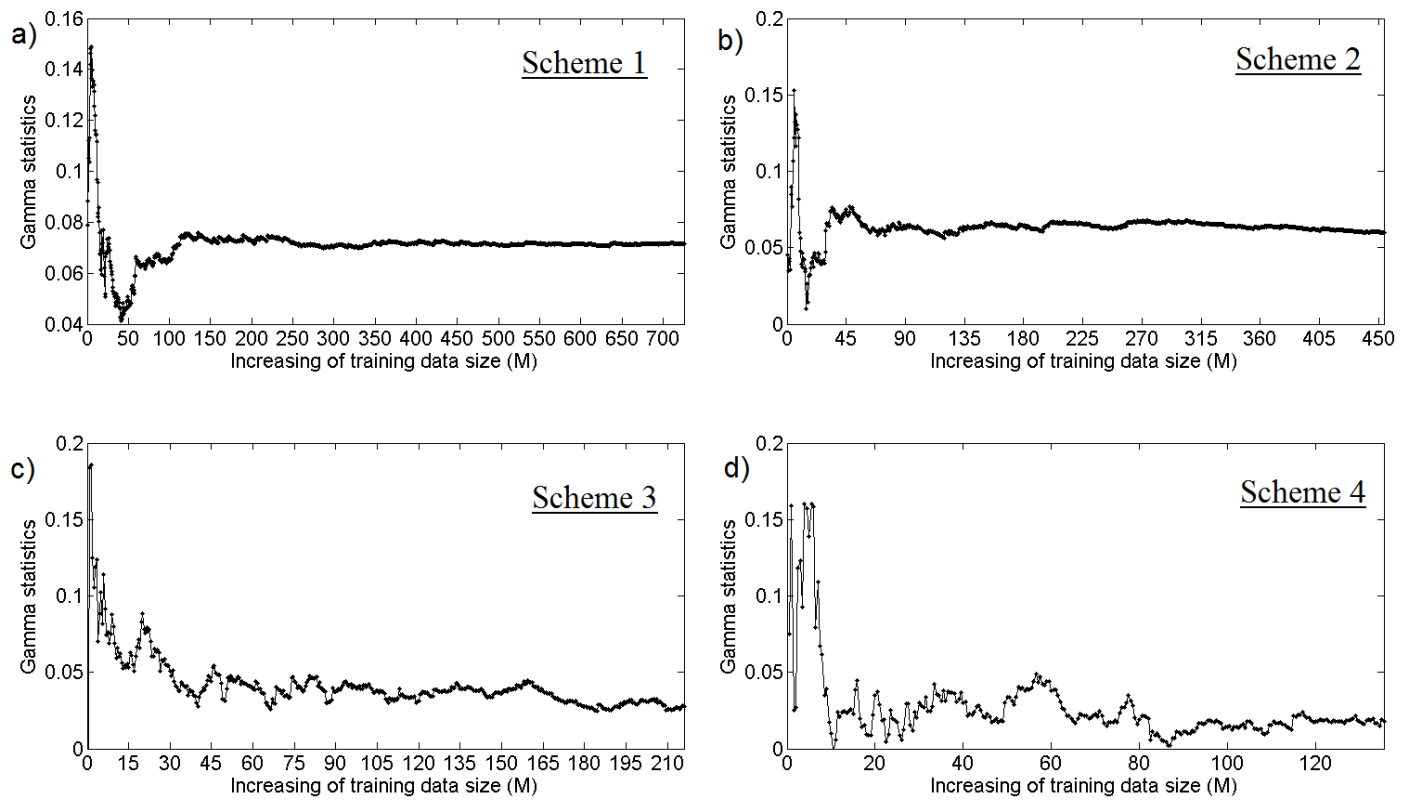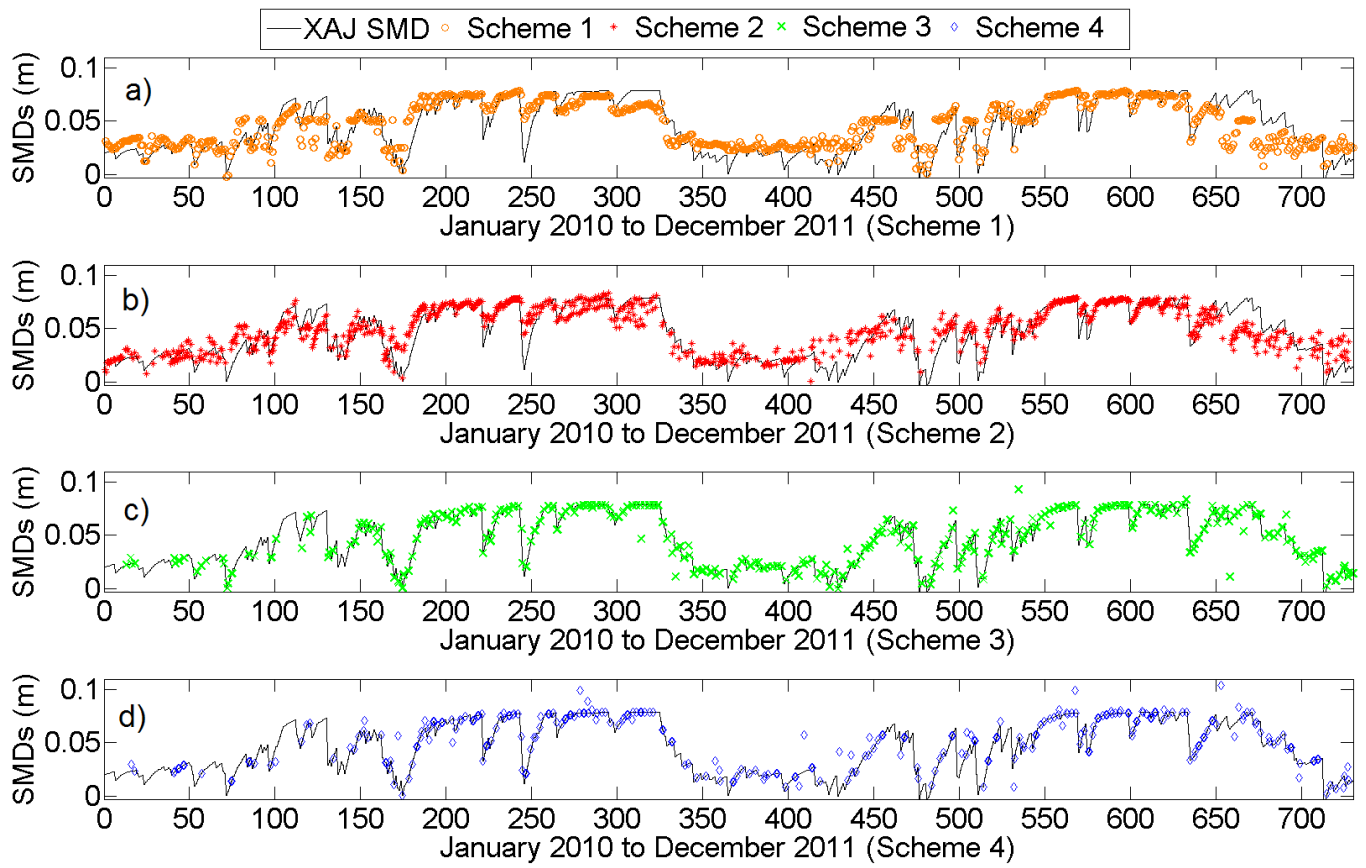
Figure 4. *M*-test, to find the best training data size: a) Scheme 1; b) Scheme 2; c) Scheme 3; and d) Scheme 4.
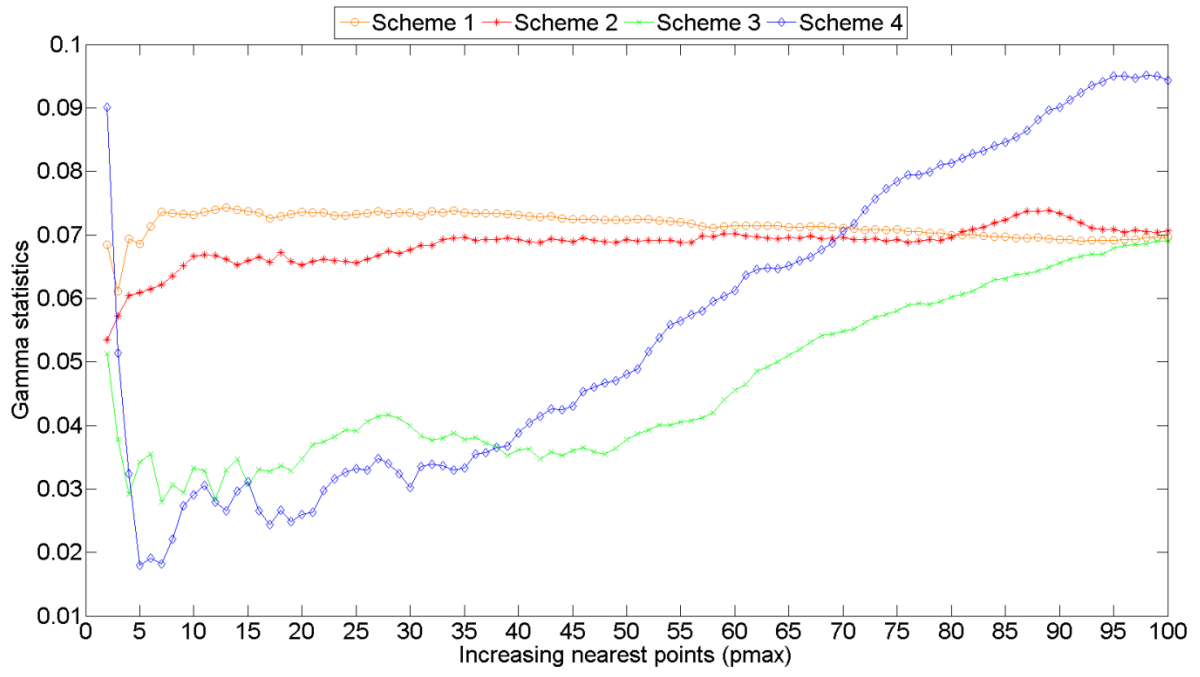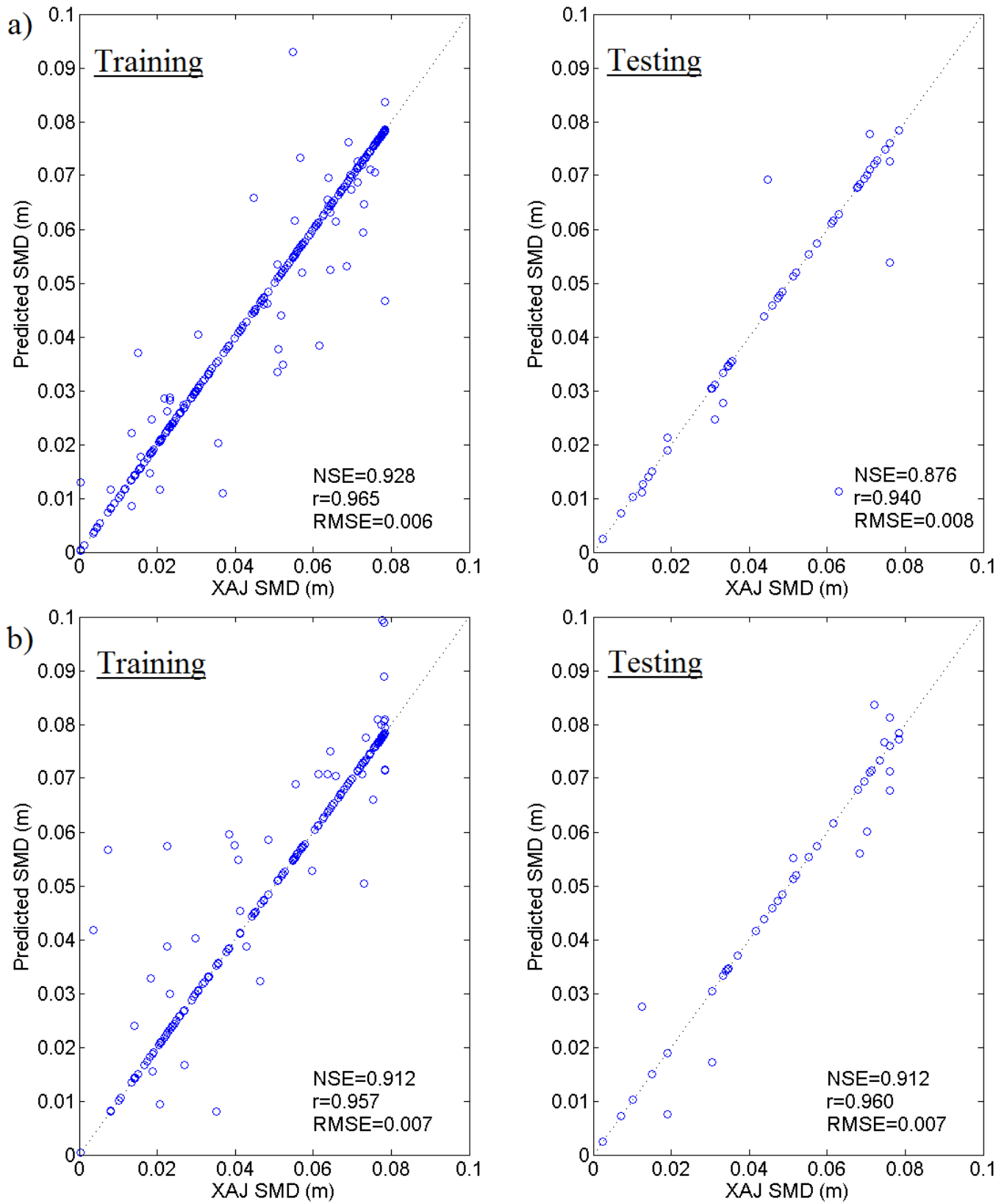
**Figure 5.** Gamma statistic (Γ) variations for increasing the LLR p_{max} value.

**Figure 6.** LLR modelling during the training and testing phases for a) Schemes 1 and b) Scheme 2.

Figure 7. The time series plots of the XAJ SMD and the estimated SMD from the four schemes: a) Scheme 1; b) Scheme 2; c) Scheme 3; and d) Scheme 4.

**Figure 8.** LLR modelling during the training and testing phases for a) Schemes 3 and b) Scheme 4.
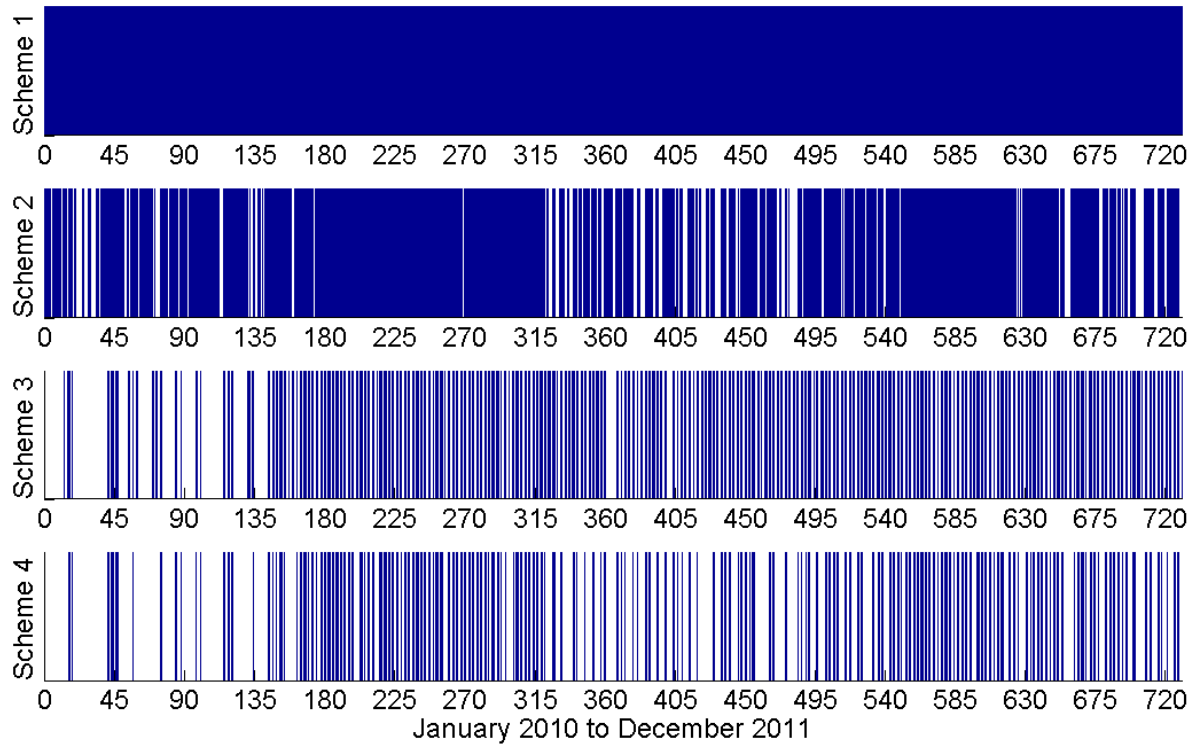
Figure 9. Data availability plots of the four schemes: Scheme 1: SAC-SMA-SM input; Scheme 2: SAC-SMA-SM and MODIS-LST inputs; Scheme 3: SAC-SMA-SM and SMOS-T_bs inputs; Scheme 4: SAC-SMA-SM, MODIS-LST, and SMOS-T_bs inputs. The total available days for the four schemes are 730, 458, 217, and 140 respectively.
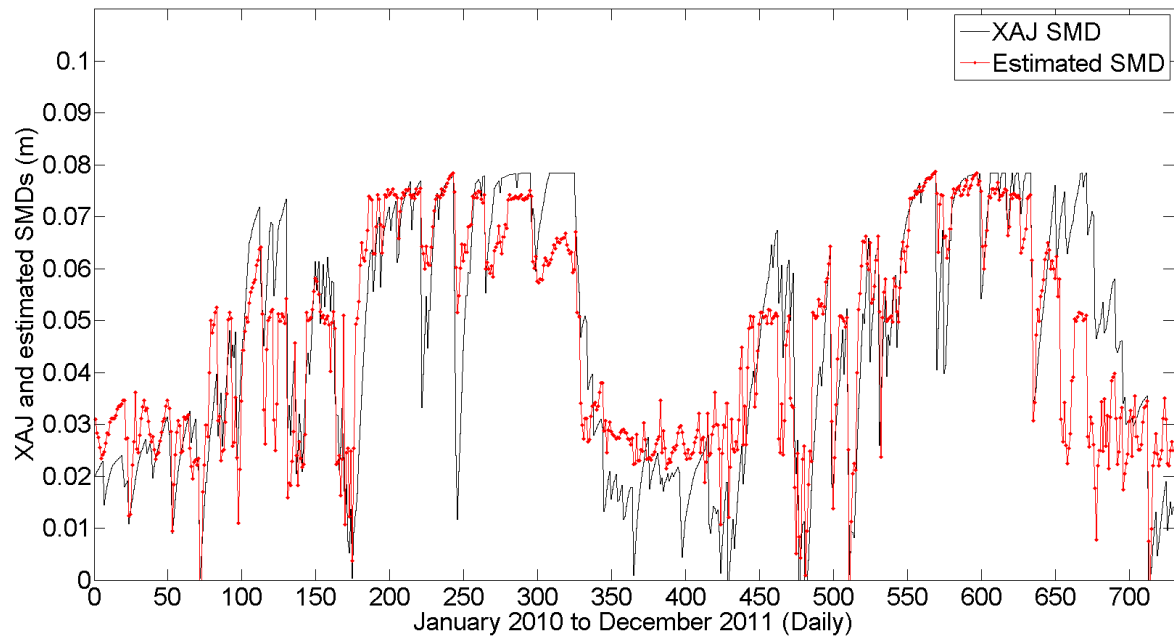
Figure 10. Time series plot of the combined daily hydrological soil moisture state estimations.
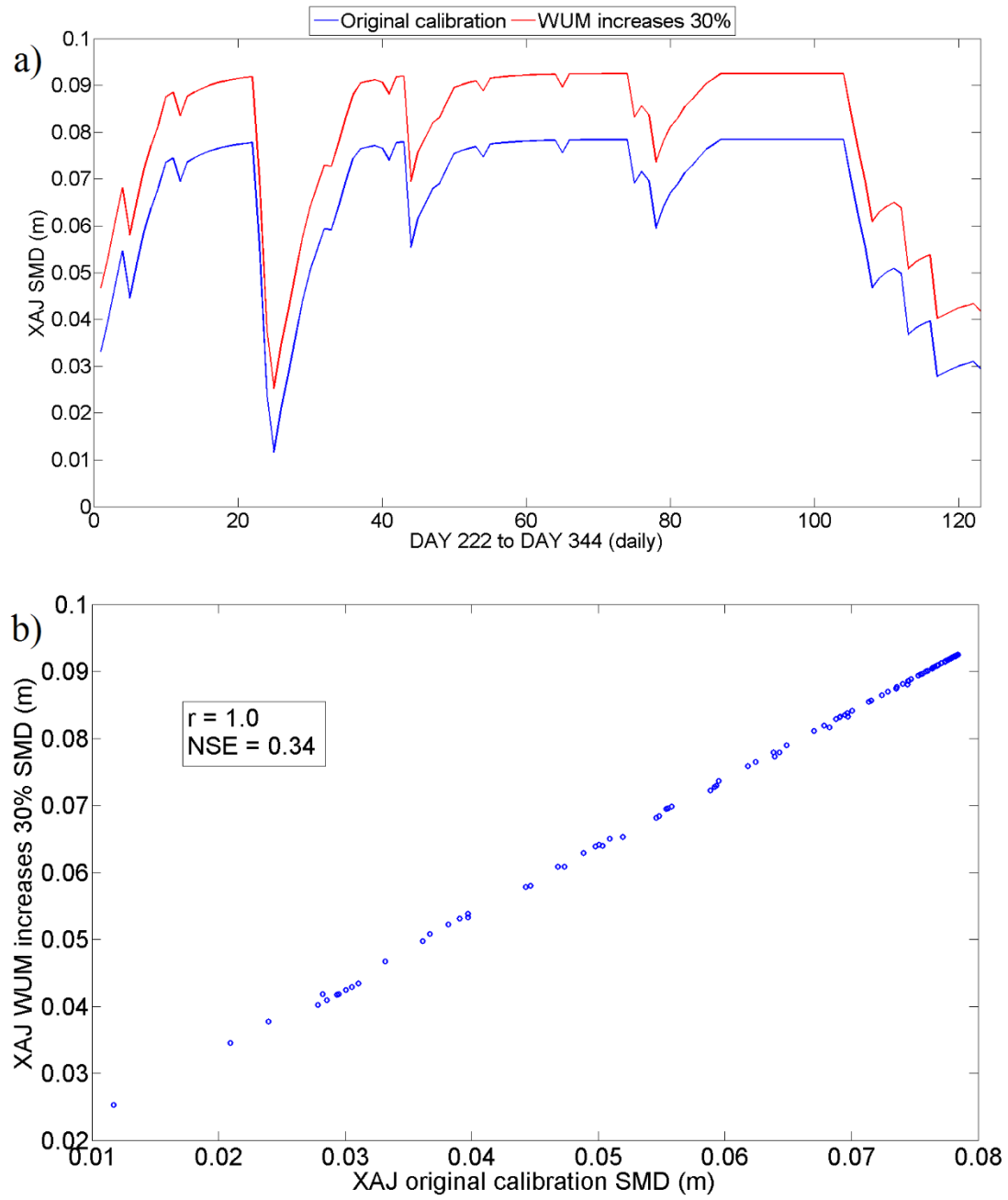
Figure 11. SMD variations from the manipulated XAJ calibration (i.e., the WUM parameter is increased by 30 %) and its original calibration.
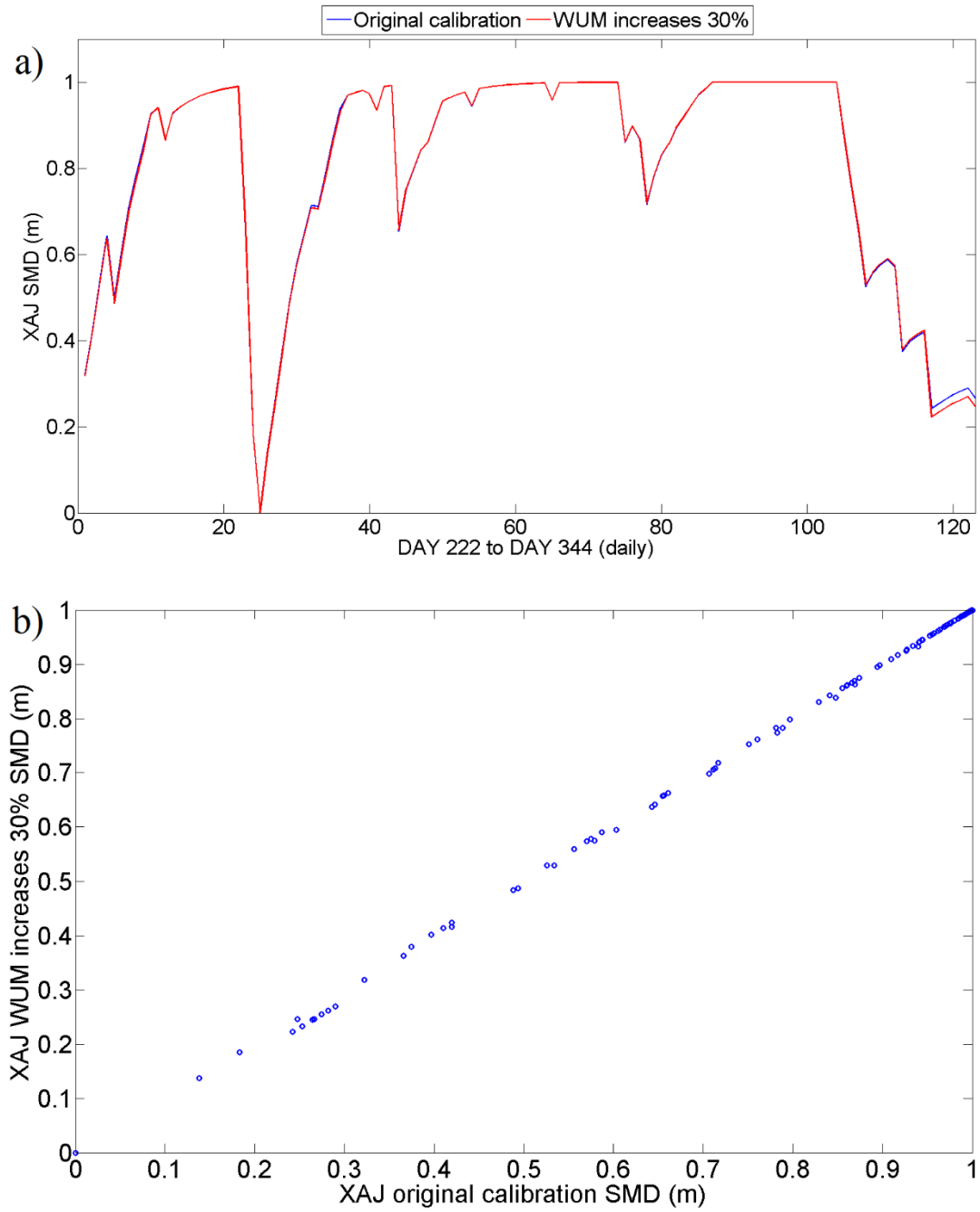
Figure 12. Normalised SMD variations from the manipulated XAJ calibration (i.e., the WUM parameter is increased by 30 %) and its original calibration.