To Editor

Dear Prof. Gelfan,

Thank you very much for accepting the paper with minor revision. The manuscript has been amended to address the issues raised in the reviews. The point-by-point replies to the referees' comments are listed below.

If there are any further questions, please let us know.

Lu Zhuo

The Corresponding author (lu.zhuo@bristol.ac.uk)

University of Bristol

Replies to Referee #1

....

General comments. The paper is devoted to investigation of ability to estimate land surface soil moisture with the help of the Local Linear Regression model using data retrieved from hydrological model SAC-SMA (soil moisture), radiometer MODIS (land surface temperature), and SMOS (brightness temperatures in H-V polarizations). The research topic is very relevant. The described approach is original and the results is quite convincing.

Specific comments. These comments concern with literature redaction of the paper. May be the Introduction is too long and its text is partly repeated in different sections of the paper. The text is quite loose and without prejudice to the content may be compressed in different parts of the paper especially in the beginnings of the sections.

Reply: Agreed. Some contents have been moved to Section 2 Material and Methods, so the revised Introduction is now more condensed.

Technical corrections 1. It is necessary to correct "polarisation" for "polarization" (for instance, line 13 and further). This mistake can be corrected using contextual replacement.

Reply: They have been corrected.

There are some faults in the References.

Line 602. Is winGammaTM the beginning of PhD Thesis title?

Reply: Yes. It is the beginning of the PhD Thesis title.

Line 642. smos mission has be replaced by SMOS mission

Reply: It has been corrected.

Lines 656-657. JOURNAL OF GEOPHYSICAL RESEARCH-ALL SERIES-, 99: 14,415-14,415. The letters must be in lower case. And what about pages?

Reply: The reference has been updated.

Line 701. Number of pages is not specified

Reply: The page number has been added.

Lines 780-783 One paper is mentioned instead of two different ones. 780 Zhao, R.-J., 1992a. The Xinanjiang model applied in China. Journal of Hydrology, 135(1): 781 371-381. 782 Zhao, RenJun, 1992b. The Xinanjiang model applied in China. Journal of Hydrology, 135(1): 783 371-381

Reply: They have now been combined as one reference.

Resume The submitted paper is interesting; it corresponds to the subject of the journal. The paper may be published in the HESS after mentioned corrections.

....

Replies to Referee #2

The authors have suggested a method for creating a soil moisture product for hydrological applications using multiple data sources retrieved from three sources (SAC-SMA land surface modelling product, MODIS satellite-retrieved land-surface temperature, and Soil Moisture and Ocean Salinity (SMOS) project data) using the Gamma-test and the Local Linear Regression techniques. The accuracy of the produced soil moisture data was evaluated against the Xinanjiang (XAJ) hydrological model's soil moisture simulations. The authors have concluded that "together with the chosen data inputs can be used with high confidence to estimate an unintermitted hydrological soil moisture product, and the proposed method could be easily applied to other catchments and fields". The topic is of current scientific interest and the manuscript is overall well prepared. However, there are some general points that need to be clarified and at some points more detailed information or analysis is necessary. The following general and specific comments should be addressed before accepting this manuscript for publication.

General Comments

1. Duration of the study period is two years: from January 2010 to December 2011. In my opinion, the presented calibration/validation results don't allow a reader to evaluate the model applicability and the aforementioned conclusion looks too optimistic. The point is that the study period is too short, and the presented results of the model validation are deficient. It means that

the overall model performance based on these results is very sensitive to the meteorological conditions of the study period and the performance assessments are rather casual. This fact limits opportunities for application of the proposed method "to other catchments and fields". The conclusion on the model applicability would be more convincing if the authors evaluated the model against hydrological data for a longer period. According to the USGS Water Resources webpage, streamflow data series exceed 10 years for the Vermilion River at Pontiac, IL.

Reply: We agree with the referee's suggestion on using a longer period for model evaluation. However this study has been constrained due to the following two reasons: first although the selected catchment has flow data over 10 years, most of them are discontinuous (e.g., frequent data gaps), and the selected period provides the most complete flow observations. This is essential for the XAJ model's calibration and validation (it is a continuous simulation model instead of an event-based model); second the SMOS satellite was launched in late 2009, so its data products are only available since then. In addition, since this study is a proof of concept for a new method in optimal fusing of multi-source data for soil moisture estimation, it is more effective to inform the hydrological community about such an approach and trial it in a variety of locations than testing it at one location for a few more years.

2. The Pontiac catchment is characterized by frequent soil freezing events in winter seasons. During freezing events, soil moisture transfer fundamentally differs from the unfrozen conditions (e.g. Gelfan, 2006). To my knowledge, the lumped XAJ model does not consider soil freezing, thus SMD simulations can be inaccurate for winter seasons. Please clarify.

Gelfan A. N. (2006) Physically based model of heat and water transfer in frozen soil and its parametrization by basic soil data. IAHS Publ., 303, pp. 293-304.

Reply: The lumped XAJ model has been frequently used in frozen soil conditions (e.g., see Application of Xin'anjiang model in severe cold region of Niqiu River, 2008 (Zhou et al., 2008)). The XAJ model's evapotranspiration component plays a vital role in the model's flow generation. The component constitutes three soil storages as seen in Fig.2 (i.e., the SMD simulations). The accuracy of SMD can therefore largely affect the accuracy of the model's flow calculation. As seen in Fig.3 (in original manuscript), XAJ model is very good at simulating the flow variations in the Pontiac catchment with high NSE values, even during the winter season, which agrees well with the literature. This indicates the model's SMD simulations are quite realistic.

Zhou, S., Li, Y., Zhu, J., 2008. Application of Xin'anjiang model in severe cold region of Niqiu River. Water Resources & Hydropower of Northeast China, 26(9). DOI:10.3969/j.issn.1002-0624.2008.09.016

3. The authors argued that "only the surface SMD referring to the vegetation and the very thin topsoil, is utilised as a hydrological soil moisture target". Does the XAJ model allow one to

simulate SMD in the "very thin topsoil"? If no, this should be clearly pointed out in the manuscript, and the simulation results' interpretation should be corrected.

Reply: XAJ model has three soil layers in its modelling structure, and yes the top layer represents the very thin topsoil. We have made this clearer in the updated manuscript.

4. I fully agree with the authors that the results are "model parameter dependent" (line 486). But I disagree that the proposed NHSMS indicators allow one to obtain independent results. I think that Figure 13 cannot be considered as an evidence of such independence because of at least two reasons: (1) only one parameter has been changed; (2) the obtained closeness of the two curves is shown for only 4 months (of 2 years) with mostly high SMD values. Thus I believe that the results presented in the manuscript are dependent on the XAJ model structure, parameters and inputs. Please give a comment.

Reply: It is true the proposed NHSMS only considers the model parameter factor, and other factors such as model structure is required in our future studies. As a result, the proposed indicator is only the very first step into creating a universal soil moisture product, and a lot of case studies adopting different hydrological models will need to be carried out. This has been pointed out in the original manuscript Lines 493-500 ("In the future it is planned to use the same process on other hydrological models to test if the normalised soil moisture indicators are not only model parameter independent but also model structure independent. Since all hydrological models are driven by the same hydrological inputs (precipitation, evapotranspiration and flow), their normalised soil moisture indicators should respond in a similar way (soil becomes wetter when it rains and drier when there is no rain). If this is true a new soil moisture product based on NHSMS could be generated as a routine product by the operational organisations such as NASA and ESA.").

The reason of showing the 4 months results is for a better visualisation. The selection of a dry period (i.e., high SMD values) is because it is the most critical period of time for the need of accurate soil moisture values for hydrological modelling. A good analogy is a hydrological model's water storage can be seen as a tea cup, during wet season with a lot of rain, it is easy to calculate how much water comes out of the cup as we know the maximum storage amount (for a hydrological model, it is via parameter calibration), so it is the total rainfall minus the maximum storage. However during dry season, it is impossible to know the outflow amount.

Therefore during the real time flood forecasting, after a long period of dryness, the accumulation of error in the hydrological models can become larger and larger with time. With accurate soil moisture information, the error could be corrected. This clarification has been added in the updated manuscript.

5. Conclusion, lines 539-541. I do not share the optimistic view of the authors on the perspective of the proposed fusion technique. Yes, the data sources contain part of useful (and probably independent) information. However, these data contain their own large measurement

errors and error's synergy can result in dramatically increase of the presented results' uncertainty. I would like to read the authors' comment on this topic.

Reply: It is true that all data sources have their own errors and the errors' synergy could increase the uncertainty of the merged data. This is why a good data fusion scheme should be explored and adopted. In this study, when combining them together using the data fusion method, the total error is controlled by the desired target (the XAJ model's SMD, which has been regarded as quite realistic due to the model's good flow simulations results).

Specific Comments

1. Lines 133-136: The phrase beginning from the words "It is worth noting that. . ." looks unnecessary in scientific (non-popular) hydrological text

Reply: Agreed. It has been removed.

2. I suggest removing Fig. 3. This figure has been already demonstrated in three (at least) recently published papers of the authors (Zhuo and Han, 2016, 2017; Zhuo et al., 2015)

Reply: Agreed. It has been removed.

3. Line 165: There are 17 parameters in Table 1

Reply: Agreed. It has been corrected.

4. Eq. 1: M is not defined

Reply: The definition has been added.

5. Lines 495-496: The statement "all hydrological models are driven by the same hydrological inputs (precipitation, evapotranspiration and flow)" is misconception. Please be more precise.

Reply: This statement has been modified as "all hydrological models are driven by the same physics laws on the conservation of mass…"

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   ~~polarisation~~polarizations). The simple yet effective Local Linear Regression model is applied

14   for the data fusion purpose in the Pontiac catchment. Four schemes according to temporal

15   availabilities of the data sources are developed, which are pre-assessed and best selected by

16   using the well-proven feature selection algorithm Gamma Test. The hydrological accuracy of

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

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

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

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

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

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

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

24 and 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

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

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

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

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

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

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

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

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

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

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

observations by satellite sensors).

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

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

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

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

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

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

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

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

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

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

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~~b~~) 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.). ~~XAJ is the first hydrological model~~

77 ~~adopting the multi-bucket variable size method in its modelling concept which has been~~

78 ~~followed by many famous operational hydrological models (Beven, 2012), so it is~~

79 ~~representative for those similar models.~~ The reason for adopting XAJ is explained in the

80 following section. For the purpose of hydrological soil moisture state estimation, it is effective

81 to adopt the data driven method, which can map multiple data sources into the desired dataset

82 without computational burden. ~~Various data fusion techniques have been developed (Prakash~~

83 ~~et al., 2012; Srivastava et al., 2013; Wagner et al., 2012), however their methods require high~~

84 ~~computational time to run and this, in a real-time flood forecasting framework, could not match~~

85 ~~the operational needs.~~ In this study the ~~Comparatively~~ Local Linear Regression (LLR) model

86 is used. ~~is a simpler method and requires relatively low computational time. Therefore it is~~

9

87 ~~chosen in order to test if a simple method is able to provide effective performance.~~ The multiple

88 data sources applied in this study include the SMOS (Kerr et al., 2010b) multi-angle brightness

89 temperatures (T_bs) with both horizontal (H) and vertical (V) ~~polarisation~~polarizations, the

90 Moderate Resolution Imaging Spectroradiometre (MODIS) (Wan, 2008) land surface

91 temperature, and the soil moisture product by SAC-SMA (Xia et al., 2014). ~~The main reason~~

92 ~~for choosing those three data sources is due to their Near-Real-Time (NRT) availabilities~~

93 ~~(MODAPS Services, 2015; Rodell, 2016) (SMOS becomes available in NRT recently (ESA~~

94 ~~Earth Online, 2016)), which allows fast implementation in flood forecasting.~~ The detail

95 explanations of those datasets are covered in the methodology section. A well-proven feature

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

97 to pre-assess the selected data inputs and find the optimal combination of them for soil moisture

98 state calculation. In addition, an M-test (Remesan et al., 2008) is adopted to explore the best

99 size of the training data. The desired soil moisture product is trained and tested by the XAJ

100 SMD simulation. In total four data-input schemes are developed according to the temporal

101 availability of the selected data inputs, which are then combined to give a daily hydrological

102 soil moisture product.

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

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

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

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

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

2. Material and Methods

2.1 Study Area

In this study, the Pontiac catchment (1,500 km^2, Figure 1) is used for the calibration and the validation of the XAJ model. Pontiac (40.878°N, 88.636°W) lies on the north-flowing Vermilion River, which is a tributary of the Illinois River of the state of Illinois, U.S. The worst flood in this area occurred on December 4, 1982, cresting at 5.84 m above mean sea level (MSL); and the most recent flood occurred on January 9, 2008, cresting at 5.75 m MSL, so this catchment is likely located within a winter-flooding region. Pontiac is covered with moderate canopy (the annual mean Normalized Difference Vegetation Index retrieved from the MODIS satellite is around 0.4), when compared with a densely vegetated catchment, it has more accurate soil moisture estimations from satellites (Al-Bitar et al., 2012). Based on the Köppen-Geiger climate classification, this medium sized catchment is dominated mainly by hot summer continental climate (Peel et al., 2007).With reference to the University of Maryland Department Global Land Cover Classification, it is used primarily for agriculture purpose (Bartholomé and Belward, 2005; Hansen, 1998). The soil mostly consists of Mollisols, which has deep and high organic matter, and the nutrient-enriched surface soil is typically between 60-80 cm in depth (Webb et al., 2000). The study period is from January 2010 to December 2011. The reason for using this two-year period of data is due to the discontinuity of the flow records in this catchment, and the selected period provides the most complete flow observations.

The North American Land Data Assimilation System 2 (NLDAS-2) (Mitchell et al., 2004) provides precipitation and potential evapotranspiration information to run the XAJ model. Both

129 data forces are at 0.125º spatial resolution and have been converted to daily temporal resolution.

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

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

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

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

134 evapotranspiration is approximately 1670 mm. ~~It is worth noting that the actual~~

135 ~~evapotranspiration is much less than the potential amount, because dryer soil reduces the actual~~

136 ~~evapotranspiration, and if the soil is totally dry the actual evapotranspiration will be zero~~

137 ~~regardless how large the potential evapotranspiration is.~~ The daily observed flow data are

138 acquired from the U.S. Geological Survey.

139 **2.2 Hydrological Model**

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

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

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

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

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

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

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

147 applications.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

166   referred to (Zhuo and Han, 2016b) for more details. All the XAJ's ~~16~~ 17 parameters are used

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

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

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

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

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

172 respectively) of the XAJ model ~~are shown in Figure 3~~can be found in (Zhuo et al., 2015a).

173 Discussion regarding the river flow and SMD simulation results in this catchment have been

174 published in (Zhuo and Han, 2016b), with Nash-Sutcliffe Efficiency (*NSE*) obtained larger than

175 0.80 during both the calibration and validation periods. The results are not repeated here.

176 **2.3 Multiple Data Sources for Hydrological Soil Moisture State Estimation**

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

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

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

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

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

182 allows fast implementation in flood forecasting.

183 **2.3.1 SMOS Multi-angle Brightness Temperatures (SMOS-T_bs)**

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

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

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

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

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

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

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

191 thermal emission from the Earth in both H and V ~~polarisation~~polarizations with a wide ranges

14

192   of incidence angles from 0° to 60°. The observation depth of SMOS is approximately 5 cm

193   with a spatial resolution of 35-50 km depending on the incident angle and the deviation from

194   the satellite ground track (Kerr et al., 2012; Kerr et al., 2010a; 2001).

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

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

197   ~~polarisation~~polarizations). Therefore the number of the SMOS-T_bs inputs for the hydrological

198   soil moisture estimation can be as high as 24 (12 angle bins per ~~polarisation~~polarization), with

199   the centre of the first angle bin at 2.5^o in both ~~polarisation~~polarizations (Rodriguez-Fernandez

200   et al., 2014). As satellite progresses, any given location on the Earth's surface is scanned a

201   number of times at various incidence angles, depending on the location with respect to the

202   satellite subtrack: the further away, the fewer the angular acquisitions (Kerr et al., 2010b). The

203   data availabilities of the SMOS-T_bs are illustrated in Figure ~~4~~3 (the availabilities for H and V

204   ~~polarisation~~polarizations are the same). It can be seen that the data availabilities among various

205   incidence angles are rather different. In this study the only angle range that gives the most

206   available record of data is from 27.5^o to 57.5^o (i.e., 7 for H and 7 for V ~~polarisation~~polarization),

207   which is therefore chosen for the hydrological soil moisture development. This angle range is

208   in line with the angle selection in (Rodriguez-Fernandez et al., 2014). In addition the SMOS

209   Level-3 soil moisture product from the CATDS (SMOS-SM) is also acquired for a comparison

210   with the estimated soil moisture product. Retrievals that are potentially contaminated with

211   Radio Frequency Interference have been removed. Readers are referred to (Kerr et al., 2012)

212   for a full description of the SMOS retrieving algorithms, and (Njoku and Entekhabi, 1996) for

213   a good knowledge of how passive microwave relates to soil moisture variations.

2.3.2 MODIS Land Surface Temperature (MODIS-LST)

The MODIS/Terra (Earth Observing System AM-1 platform) (Wan, 2008) daily MOD11C1-V5 land surface temperature covering the studying period is downloaded from the Land Processes Distributed Active Archive Centre website. MODIS is chosen among other operational optical satellites for its suitable features, mostly, due to its frequent revisiting time and free NRT data availability. It measures 36 spectral bands between 0.405 and 14.385 μm, and acquires data at three spatial resolutions 250 m, 500 m, and 1,000 m respectively while the adopted MOD11C1 V5 product incorporates 0.05° (5.6 km) spatial resolution. The benefit of adding land surface temperature information is that previous studies have shown the variations in soil moisture have a strong linkage with land surface temperature (Carlson, 2007; Goward et al., 2002; Mallick et al., 2009). One reason is the changes of land surface temperature are mainly affected by albedo and diurnal heat capacity, and the diurnal heat capacity is mainly controlled by soil moisture (Price, 1980). (Wan, 2008) compared MOD11C1-V5 land surface temperatures in 47 clear-sky cases with in situ measurement and revealed that the accuracy was better than 1 K in the range from −10° to 58 °C in about 39 cases. Cloud-contaminated data have been removed by a double-screening method, and its detail can be found in (Wan et al., 2002).

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

The reason for choosing the SAC-SMA land surface modelled soil moisture product is because satellite can often have missing data due to various weather and canopy conditions (e.g., rainfall, frozen weather, and vegetation coverage), so this daily dataset is essential in producing a

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

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

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

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

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

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

241   full description of the SAC-SMA data products.

242   **2.3.4 Data Availabilities**

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

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

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

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

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

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

249   2015a). Based on the data availability analysis, the proposed hydrological soil moisture product

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

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

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

253   included in all of the schemes.

254   **2.4 Data Fusion**

255   **2.4.1   Gamma Test (GT) for Feature Selection**

17

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

279   smooth functions.

280   The Γ is related to $N[i,k]$, which represents as the *k*th ($1 \leq k \leq p$) nearest neighbours of each

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

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

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

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

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

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

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

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

289   following equation:

290   $$\gamma = A\delta + \Gamma \qquad\qquad\qquad (5)$$

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

292   $$\gamma_M(k) \rightarrow Var(r), \text{ when } \delta_M(k) \rightarrow 0 \qquad\qquad (6)$$

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

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

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

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

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

298 published in (Evans and Jones, 2002).

2.4.2 *M*-test for Training Data Size Selection

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

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

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

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

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

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

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

307 problems and reducing unsystematic attempts.

2.4.3 Local Linear Regression (LLR)

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

310 2013; Wagner et al., 2012), however their methods require high computational time to run and

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

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

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

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

315 2011; Pinson et al., 2008; Sun et al., 2003; Zhuo et al., 2016b) for forecasting and smoothing

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

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

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

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

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

321    $Xm = y$ (7)

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

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

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

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

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

326   (8)

327   In order to solve the equation, the following two conditions are set: a) if X is square and non-

328   singular then Eq. (7) can be simply calculated as $m = X^{-1}y$; b) if X is not square or singular,

329   Eq. (7) needs to be rearranged and m can be get by finding the minimum of:

330    $\left| Xm - y \right|^2$ (9)

331   with the distinct solution of:

332    $m = X^{\#}y$ (10)

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

334   **3. Results**

335 In this section, different combinations of input data (Table 3) are adopted to examine their

336 impacts on hydrological soil moisture estimation. XAJ SMD is used as a hydrological soil

337 moisture state benchmark for the training and testing. More discussion about the misconception

338 between the hydrological model's soil moisture state variable and the real-world soil moisture

339 content is covered in Section 4. During GT and M-test processes, all data inputs need to be

340 normalised so that their mean is zero and standard deviation is 0.5. This step is necessary in

341 reducing the impacts of numerical difference from various inputs, hence improves the GT

342 efficiency (Remesan et al., 2008). Five statistical indicators are used for the soil moisture

343 estimation analysis: Pearson product moment correlation coefficient (r), MSE which is the

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

345 Root Mean Square Error ($RMSE$).

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

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

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

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

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

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

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

353 size. As presented in Figure 5a4a, when more training data (i.e., M increases in steps of one) is

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

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

22

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

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

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

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

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

361   its complexity level is controlled by the p_{max} parameter. As illustrated in Figure 6 5, p_{max} is

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

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

364   Figure 6 5 that the smallest Γ is achieved at p_{max} = 4, which is therefore adopted for the LLR

365   modelling. The training and testing scatter plots for the LLR modelling are shown in Figure

366   7a 6a. It is observed that there are some points lying far above the bisector line during the

367   training period signifies higher estimations whereas some points sit far below the bisector line

368   during the testing period indicates under-estimation of the SMD. For the testing results, when

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

370   at around 0.080 m) the predicted soil moisture state is still in the drying progress. Figure 8a 7a

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

372   SMD follows the seasonal trend of the soil moisture fluctuations well, so it is wetter during the

373   winter season and exsiccated during the hot summer season. However it is clear to see that the

374   model is not able to capture the extreme situations very well, especially during the wet season

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

376   **3.2 Scheme 2: SMD Estimation Using SAC-SMA-SM and MODIS-LST as inputs**

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

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

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

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

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

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

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

384   M-test in Figure ~~5b~~ 4b shows the graph settles to an asymptote around 0.060 which is consistent

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

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

387   result in Figure ~~6~~5. The LLR training and testing performances are presented in Figure ~~7b~~6b.

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

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

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

391   The improvement can also be seen clearly in the time series plot in Figure ~~8b~~7b. For example,

392   the big disparities between the estimated and the targeted SMDs around DAY 300 and DAY

393   350 are reduced evidently.

3.3 Scheme 3: SMD Estimation Using SAC-SMA-SM and SMOS-T_bs as inputs

395   The multi-angle T_bs retrievals are the main data inputs for SMOS soil moisture calculation,

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

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

24

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

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

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

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

402 2.3.4, SAC-SMA-SM is a compulsory data input, so it is not included in the selecting process.

403 The best set of SMOS-T_bs to retrieve soil moisture state is composed of H

404 ~~polarisation~~polarization at the incidence angles of 27.5°-47.5°, 57.5°, and V

405 ~~polarisation~~polarization at the incidence angles of 27.5°-42.5°, 52.5°, 57.5°. This result

406 demonstrates that using a combination of H and V T_bs gives a better soil moisture estimation,

407 which is logically sensible because different ~~polarisation~~polarizations carry distinct

408 information of the Earth surface. However some incidence angles could held common features

409 which when putting together could result in a negative effect to the LLR modelling, and are

410 therefore not included. The detailed investigation of the possible common features is out of the

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

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

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

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

415 Scheme 1's. The M-test analysis in Figure ~~5c~~4c produces an asymptotic convergence from 120

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

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

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

419 p_{max} that gives the smallest Γ is 7 (Figure ~~6~~5). The SMD estimations during the training and the

420 testing are presented in Figure 9a8a. It can be seen that the SMD prediction ability of this

421 scheme is remarkably better than the previous ones, as most of the points lie on the bisector

422 line albeit there are still some under- and over- estimations. The reason SMOS outperforms

423 MODIS in SMD estimation could be due to the long wavelength microwave has, so it presents

424 the top few centimetres of the soil while MODIS LST (thermal infrared) only provides

425 information at the soil surface. The used LLR algorithm has been double checked to filter out

426 the potential of overfitting problem. The checking processes are performed by muddling the

427 SMD target in the testing datasets as well as altering the input file, and its efficiency stays the

428 same. Hence it is believed that the LLR model is very useful in calculating SMD from this

429 scheme. Generally the *NSE*, *r* and *RMSE* statistical indicators show a high agreement during

430 both training and testing phases. For the time series plot in Figure 8c7c, it is clear to see that

431 most of the estimated points lie closely to the benchmark line. The observed outliers could be

432 partly due to the data shortage in this scheme, so that not all the scenarios are covered in the

433 datasets.

434 **3.4 Scheme 4: SMD Estimation Using SAC-SMA-SM, MODIS-LST, and SMOS-T$_b$s as**

435 **inputs**

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

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

438 suitable incidence angles from the SMOS-T$_b$s. This is because the added MODIS-LST data

439 could carry identical (i.e., redundant) features with some of the SMOS-T$_b$s datasets. As a result

440 of the full embedding calculation, the best set of SMOS-T$_b$s is composed of H

441 ~~polarisation~~polarization at the incidence angles of 37.5º-57.5º, and V ~~polarisation~~polarization

442 at the incidence angles of 37.5º-42.5º, 57.5º. As seen in Figure ~~5d~~4d, the total amount of data

443 is significanly reduced due to the shortage of simultanuously available days between the

444 MODIS and the SMOS observations. Interestingly the *M*-test graph vibrates more significantly

445 than the other three schemes, which could be due to the smaller data size and the larger amount

446 of data inputs in this scheme. Here the training data size is chosen as 62 with Γ obtained at

447 around 0.030. The optimal p_{max} is identified to be 5 (Figure ~~6~~5). The LLR modelling results are

448 shown in Figure ~~8d~~7d and Figure ~~9b~~8b. It is obvious that this scheme further improves the

449 accuracy of the SMD estimation, especially with the high statistical performances achieved

450 during both training and testing phases. Comparatively this scheme is more stable for SMD

451 estimation, albeit it requires more data inputs and is only realisable when both the MODIS and

452 the SMOS observations are available.

453 **3.5 Produce an Unintermitted Soil Moisture Product**

454 The data availability of the four schemes varies. As shown in Figure ~~10~~9, Scheme 1 which has

455 the poorest soil moisture state estimation gives the most data availability, while Scheme 4

456 which has the most accurate soil moisture state estimation owns the least data availability. In

457 order to produce an unintermitted hydrological soil moisture product, the four schemes need to

458 be combined together to complement each other. The combining method is by selecting the

459 best available soil moisture estimation. For example if all the schemes have available data at

460 the same time, the best scheme's soil moisture data is chosen (i.e., scheme 4 in this situation);

461 whereas if just one scheme has data on that day, only that scheme's soil moisture data is used.

462   The performances of the four schemes as well as the combined product are summarised in

463   Table 5. Although the combined soil moisture state is obtained with lower statistical

464   performances than Scheme 3's and 4's, it is still hydrologically very accurate especially when

465   comparing with the SMOS's official soil moisture product (Table 5). The time series of the

466   combined soil moisture state is plotted in Figure ~~11~~10. It can be seen that the general trend of

467   the produced soil moisture state follows the targeted data very well. However it tends to

468   overestimate some of the wet events during the rainy season and significantly underestimate

469   the dryer soil condition in September 2011. Those poor estimations are mostly from the Scheme

470   1 and 2 where Schemes 3 and 4 are not available. Since more and more microwave satellite

471   observations are becoming obtainable, those new data sources could add extra benefits into the

472   proposed model, and the accuracy of the soil moisture product is expected to be further

473   enhanced.

474   **4. Discussion**

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

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

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

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

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

480   each other.

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

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

28

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

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

485   simulations between Day 222 and Day 344 is presented. As seen from Figure ~~12a~~ 11a although

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

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

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

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

490   plot (Figure ~~12b~~11b) shows a very strong linear correlation (r = 1.0) between them. It is

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

492   most critical period of time for the need of accurate soil moisture values for hydrological

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

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

495   With accurate soil moisture information, the error could be corrected.

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

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

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

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

500   its connection with the real-world soil moisture.

501   -   *Soil moisture state normalisation*

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

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

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

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

506   are produced as presented in Figure ~~13~~12 (corresponding to the SMD simulations shown in

507   Figure ~~12~~11). The normalisation method is by adopting the following equation:

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

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

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

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

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

513   driven by the same physics laws on the conservation of mass~~all hydrological models are driven~~

514   ~~by the same hydrological inputs (precipitation, evapotranspiration and flow)~~, their normalised

515   soil moisture indicators should respond in a similar way (soil becomes wetter when it rains and

516   drier when there is no rain). If this is true a new soil moisture product based on NHSMS could

517   be generated as a routine product by the operational organisations such as NASA and ESA.

518   Such a soil moisture product will also be very useful to the meteorological and hydro-

519   meteorological fields in their land surface modelling because the current land surface models

520   suffer from poor performance in their runoff estimations. As aforementioned, all current soil

521   moisture products such as those from ESA and NASA are not optimised for different

522   application fields. Our study gives an example of simulating the soil moisture data targeted to

523   serve the hydrological community. It is possible other products serving farmers in agriculture,

524 ecologists in the environment, and geotechnical engineers in construction could be produced

525 using the proposed method.

526   - *Application of the produced soil moisture data*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

5. Conclusions

A hydrological soil moisture product is produced for the Pontiac catchment using the GT and the LLR modelling techniques based on four data-input schemes. Three data sources are considered including the soil moisture product from the SAC-SMA model, the land surface temperature retrieved by the MODIS satellite, and the multi-angle brightness temperatures acquired from the SMOS satellite. The four data-input schemes are built from the four combinations of the data sources. The generated soil moisture product (unintermitted with no missing data) for a period of two years (2010-2011) is compared with the XAJ hydrological model's SMD simulation to test its hydrological accuracy. It is concluded that the GT and the LLR modelling techniques together with the chosen data inputs can be used with high confidence to estimate an unintermitted hydrological soil moisture product, and the proposed method could be easily applied to other catchments and fields.

In this study it has been found that different data sources have their own unique information contents, so that they can complement each other using data fusion technique. Their synergy can be best achieved to produce an enhanced soil moisture product. In data fusion an important principle is MRmr (Maximum Relevance minimum redundancy). The soil moisture state in this study is generated from a large number of data inputs, and their selection is carried out by the GT which is one of the methods in MRmr. This is the first time that the GT is used in a data fusion of satellite multiple T_bs scans, land surface temperature and external soil moisture information for producing a hydrological soil moisture product. Future studies should explore other MRmr methods in addition to GT, to compare if they are more effective input selection

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

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

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

579 **References**

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

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

587   Bartholomé, E, Belward, AS, 2005. GLC2000: a new approach to global land cover mapping
588   from Earth observation data. International Journal of Remote Sensing, 26(9): 1959-
589   1977.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

662 Kerr, Yann H, Waldteufel, Philippe, Wigneron, J-P, Martinuzzi, J, Font, Jordi, Berger, Michael,
663 2001. Soil moisture retrieval from space: The Soil Moisture and Ocean Salinity (SMOS)
664 mission. Geoscience and Remote Sensing, IEEE Transactions on, 39(8): 1729-1735.

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

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

672   Liang, Xetal, Lettenmaier, Dennis P, Wood, Eric F, Burges, Stephen J, 1994. A simple
673   hydrologically based model of land surface water and energy fluxes for general
674   circulation models. <u>Journal of Geophysical Research: Atmospheres, 99(D7)</u>~~JOURNAL~~
675   ~~OF GEOPHYSICAL RESEARCH ALL SERIES , 99: ~~<u>:</u>14,415-14,~~415~~<u>428</u>.

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

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

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

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

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

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

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

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

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

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

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

712 Norbiato, Daniele, Borga, Marco, Degli Esposti, Silvia, Gaume, Eric, Anquetin, Sandrine,
713   2008. Flash flood warning based on rainfall thresholds and soil moisture conditions:
714   An assessment for gauged and ungauged basins. Journal of Hydrology, 362(3): 274-
715   290.

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

719 Peng, G., Leslie, L.M., Shao, Y., 2002. Environmental Modelling and Prediction. Springer.
720   480 pp.

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

723 Penrose, Roger, 1956. On best approximate solutions of linear matrix equations, Mathematical
724   Proceedings of the Cambridge Philosophical Society. Cambridge Univ Press, pp. 17-
725   19.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

779 Wan, Zhengming, Zhang, Yulin, Zhang, Qincheng, Li, Zhao-liang, 2002. Validation of the
780   land-surface temperature products retrieved from Terra Moderate Resolution Imaging
781   Spectroradiometer data. Remote sensing of Environment, 83(1): 163-180.

782 Wanders, Niko, Bierkens, Marc FP, de Jong, Steven M, de Roo, Ad, Karssenberg, Derek, 2014.
783   The benefits of using remotely sensed soil moisture in parameter identification of large-
784   scale hydrological models. Water Resources Research, 50(8): 6874-6891.

785 Wang, QJ, 1991. The genetic algorithm and its application to calibrating conceptual rainfall-
786   runoff models. Water resources research, 27(9): 2467-2471.

787 Webb, Robert W, Rosenzweig, Cynthia E, Levine, Elissa R, 2000. Global Soil Texture and
788   Derived Water-Holding Capacities (Webb et al.). Data set. Available on-line
789   [http://www. daac. ornl. gov] from Oak Ridge National Laboratory Distributed Active
790   Archive Center, Oak Ridge, Tennessee, USA.

791 Xia, Youlong, Mitchell, Kenneth, Ek, Michael, Sheffield, Justin, Cosgrove, Brian, Wood, Eric,
792   Luo, Lifeng, Alonge, Charles, Wei, Helin, Meng, Jesse, 2012. Continental-scale water
793   and energy flux analysis and validation for the North American Land Data Assimilation

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

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

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

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

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

799   ~~Zhao, R. J., 1992a. The Xinanjiang model applied in China. Journal of Hydrology, 135(1):~~

800   ~~371-381.~~

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

802   371-381.

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

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

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

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

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

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

809   satellite soil moisture measurements for hydrological applications. Hydrological

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

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

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

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

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

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

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

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

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

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

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$_b$s	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
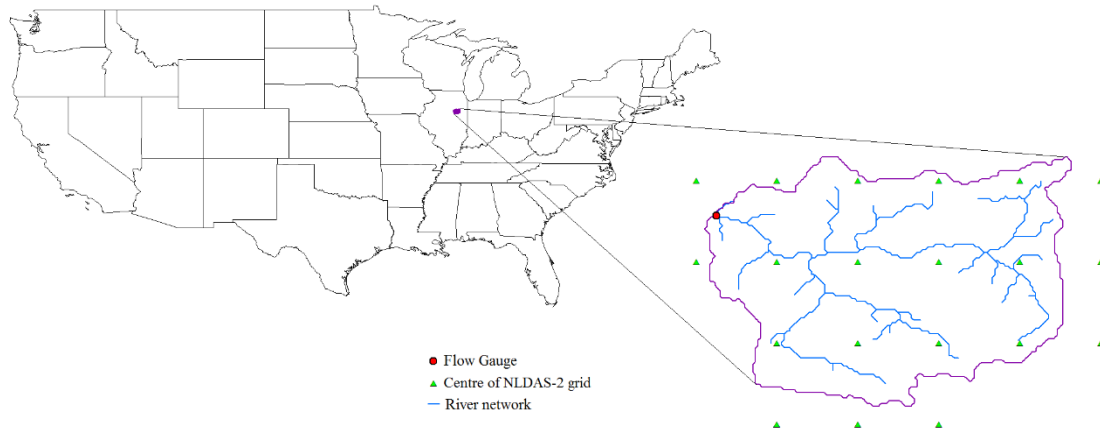
Figure. 1. The location and river network of the Pontiac catchment in the U.S., with the flow gauge and NLDAS-2 central grid points (Zhuo et al., 2015a).
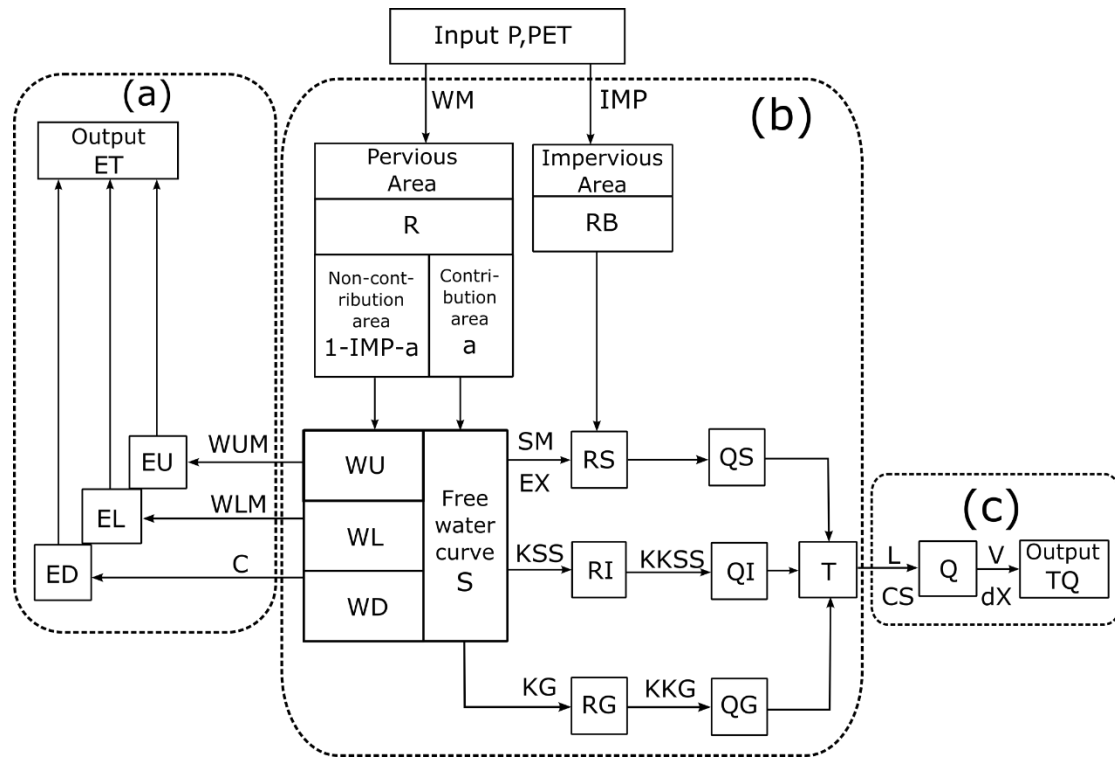
Figure. 2. Adopted flowchart of the XAJ model *(Zhao, 1992a)*. The model consists of an evapotranspiration component (a), a runoff generating component (b), and a runoff routing component (c). *P*, *PET*, and *ET* are the precipitation, potential evapotranspiration, and the simulated actual evapotranspiration 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.
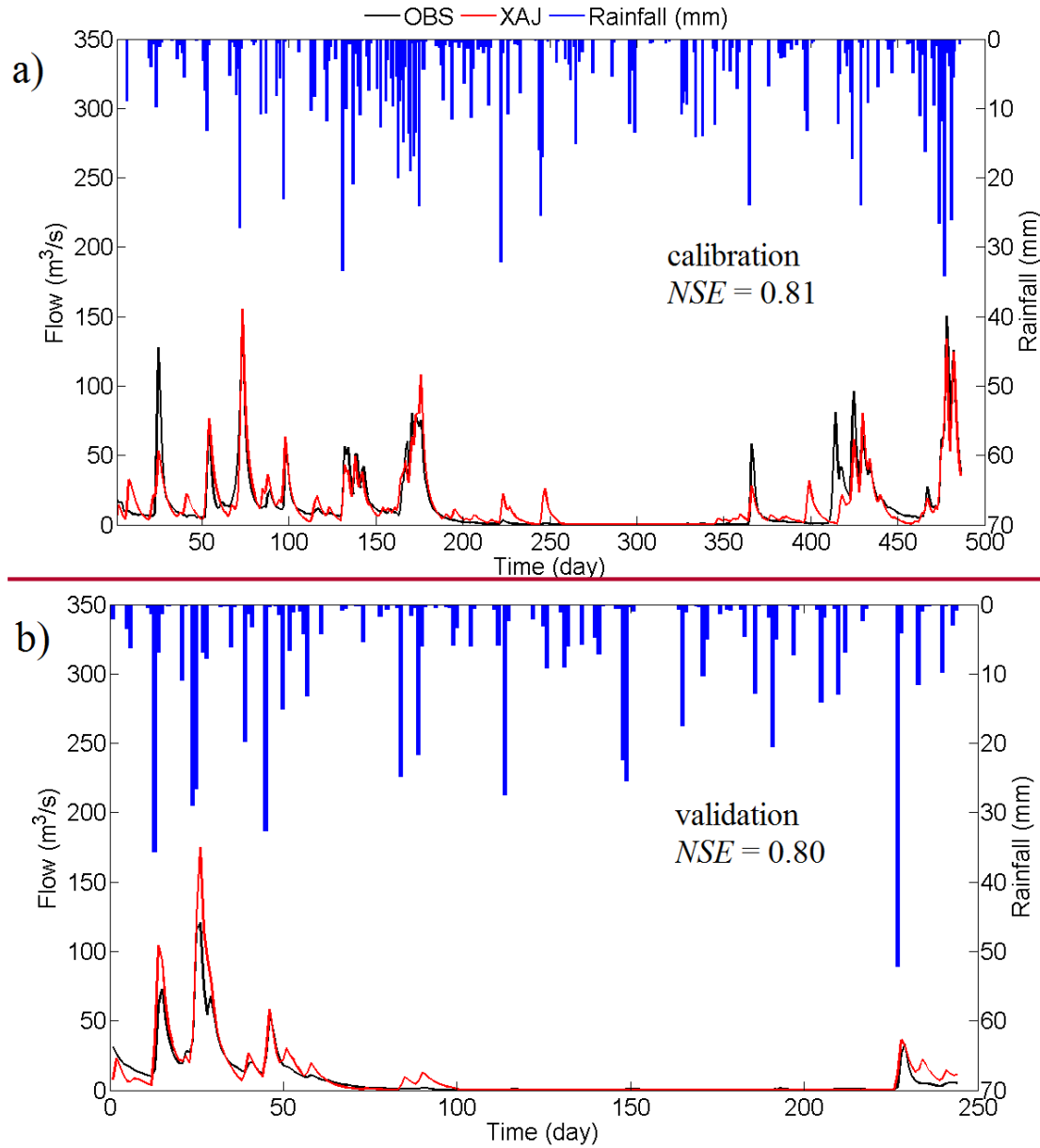
Figure. 3. Time series of daily rainfall and daily flow (observation and XAJ simulated) for the Pontiac catchment, during a) calibration and b) validation (Zhuo et al., 2015a).
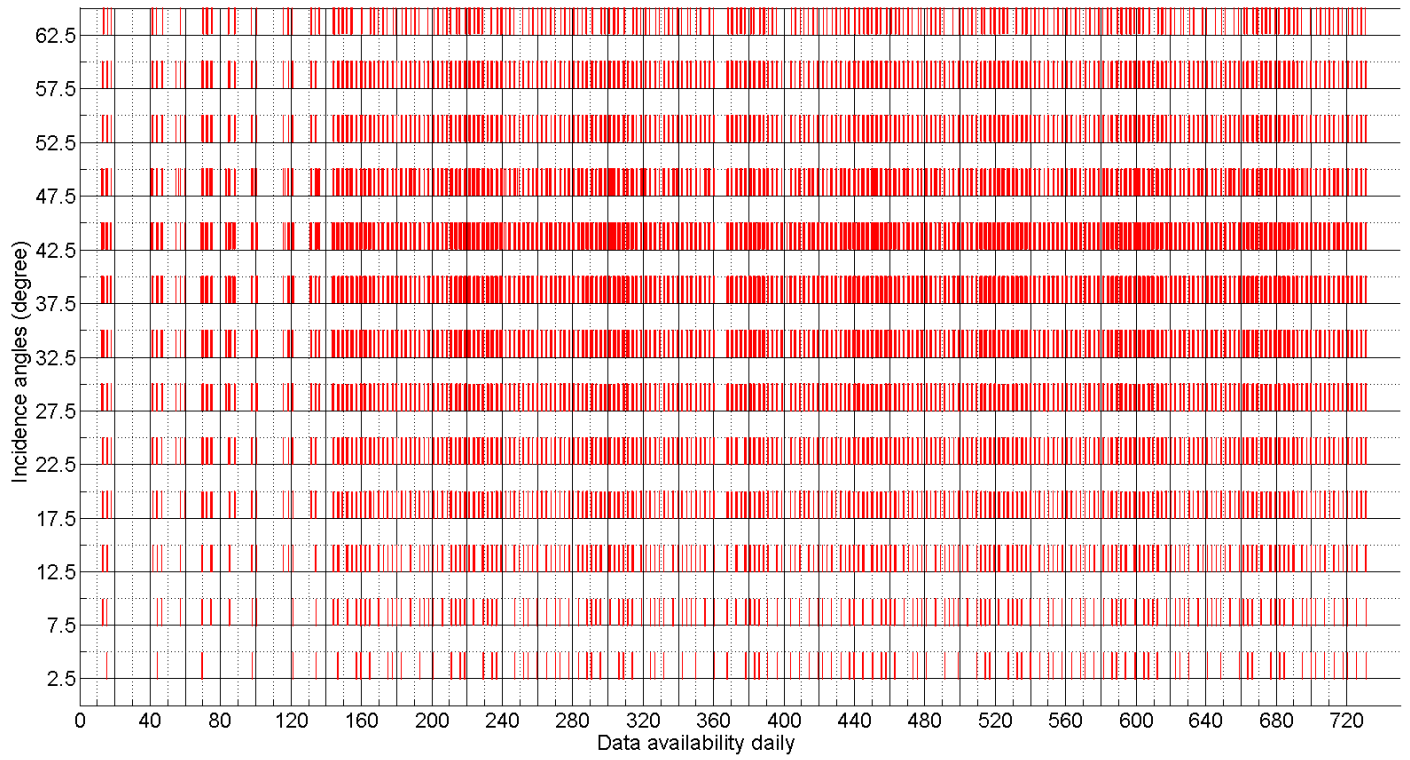
Figure 3. SMOS-T$_b$s data availabilities. It is noted that the available dates for the horizontal and the vertical ~~polarisation~~polarizations are the same, so only one is shown here.
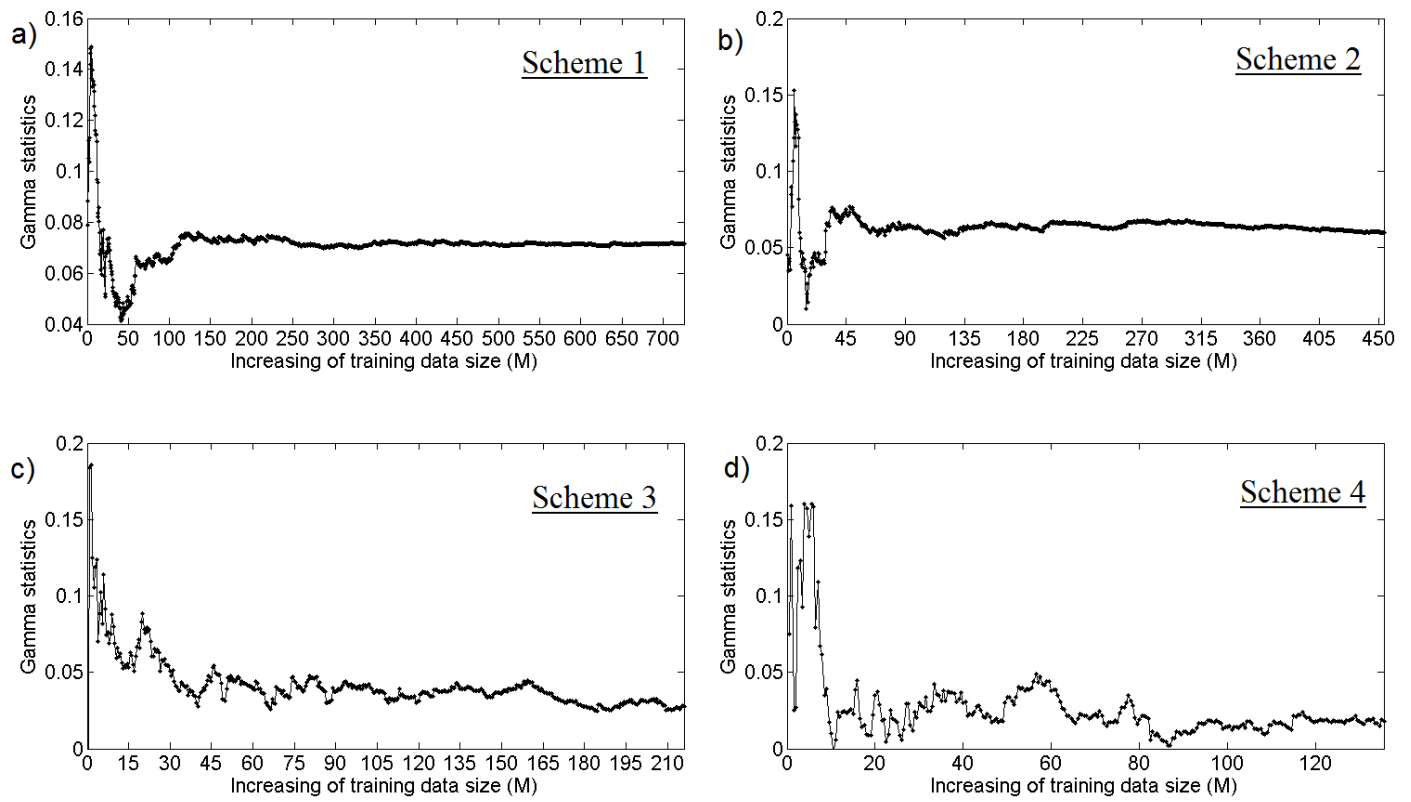
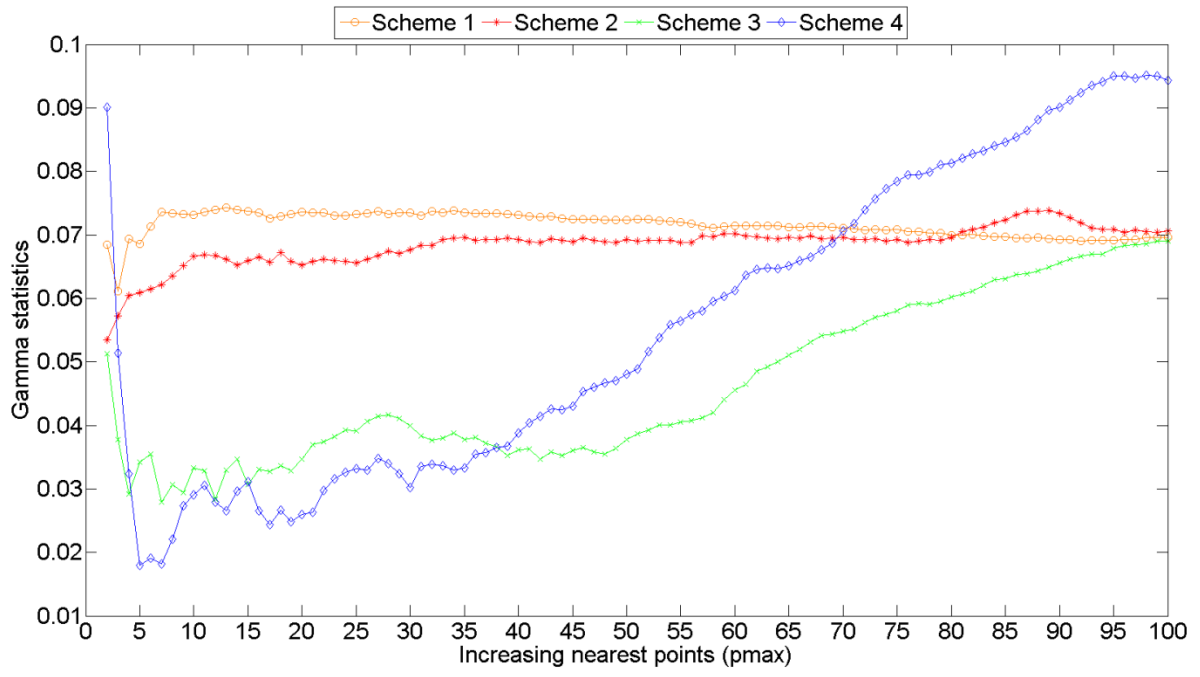**Figure 4.** *M*-test, to find the best training data size: a) Scheme 1; b) Scheme 2; c) Scheme 3; and d) Scheme 4.

Figure 5. Gamma statistic (Γ) variations for increasing the LLR p_{max} value.
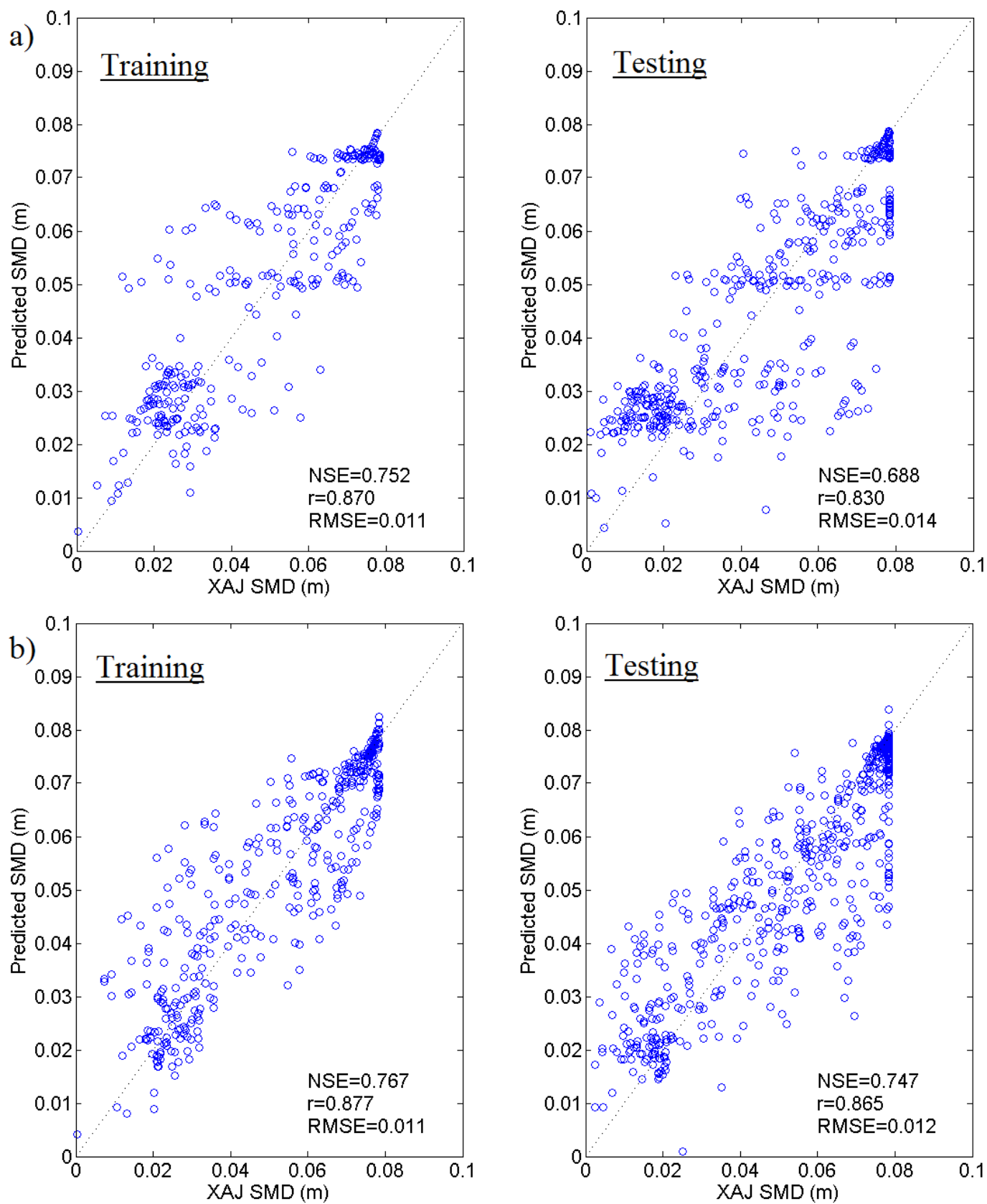
Figure 6. LLR modelling during the training and testing phases for a) Schemes 1 and b) Scheme 2.
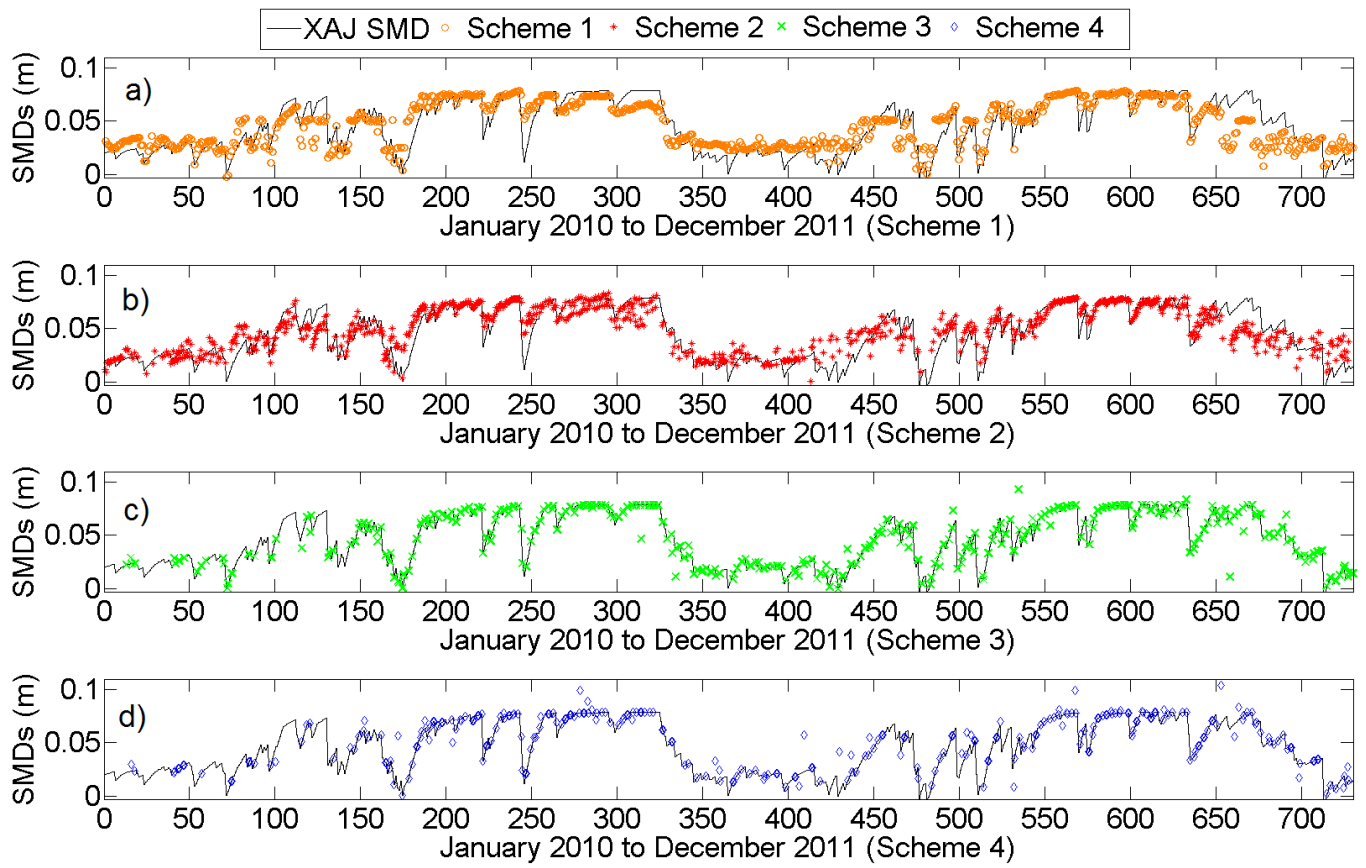
Figure 7. The time series plots of the XAJ SMD and the estimated SMD from the four schemes: a) Scheme 1; b) Scheme 2; c) Scheme 3; and d) Scheme 4.
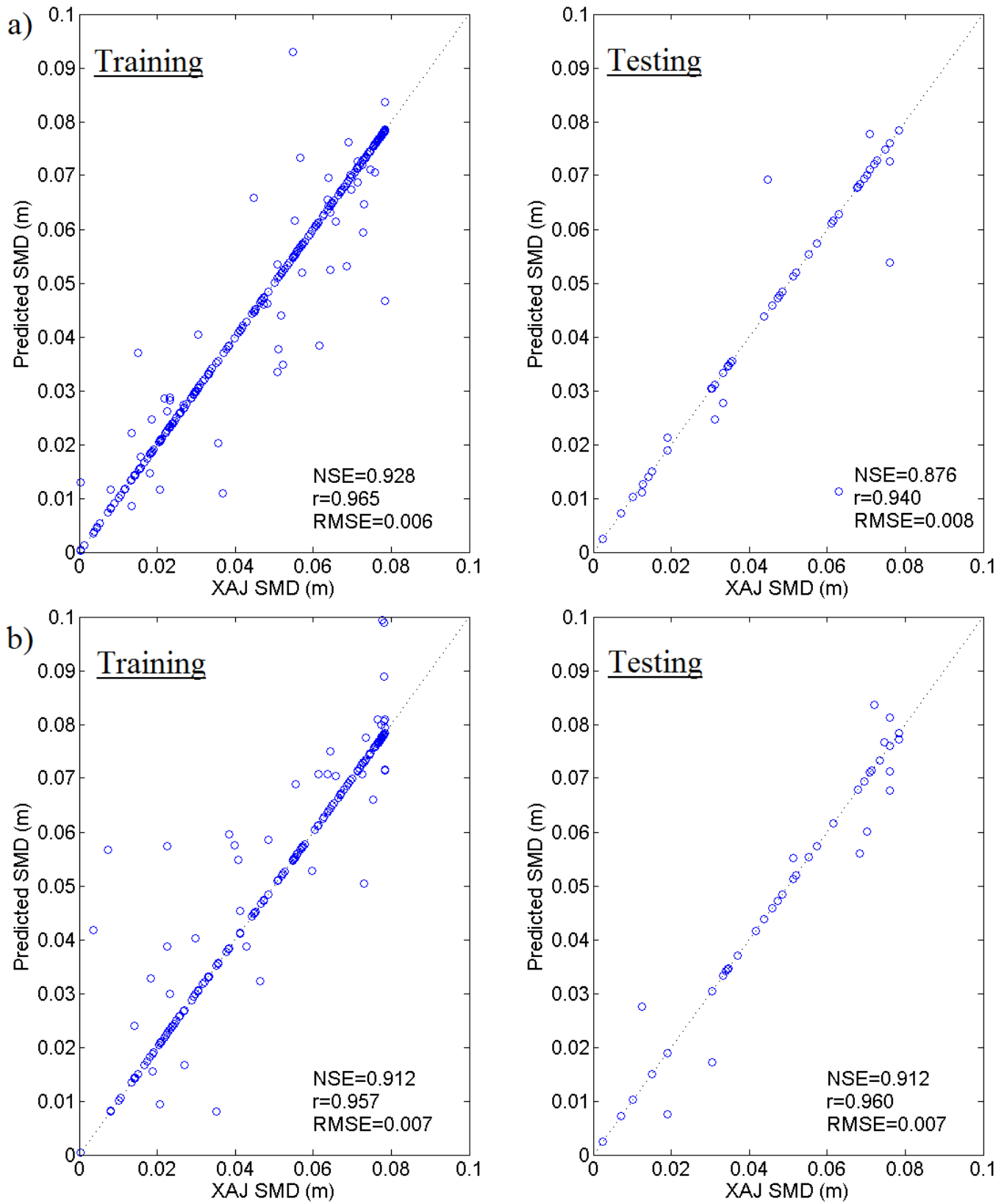
Figure 8. LLR modelling during the training and testing phases for a) Schemes 3 and b) Scheme 4.
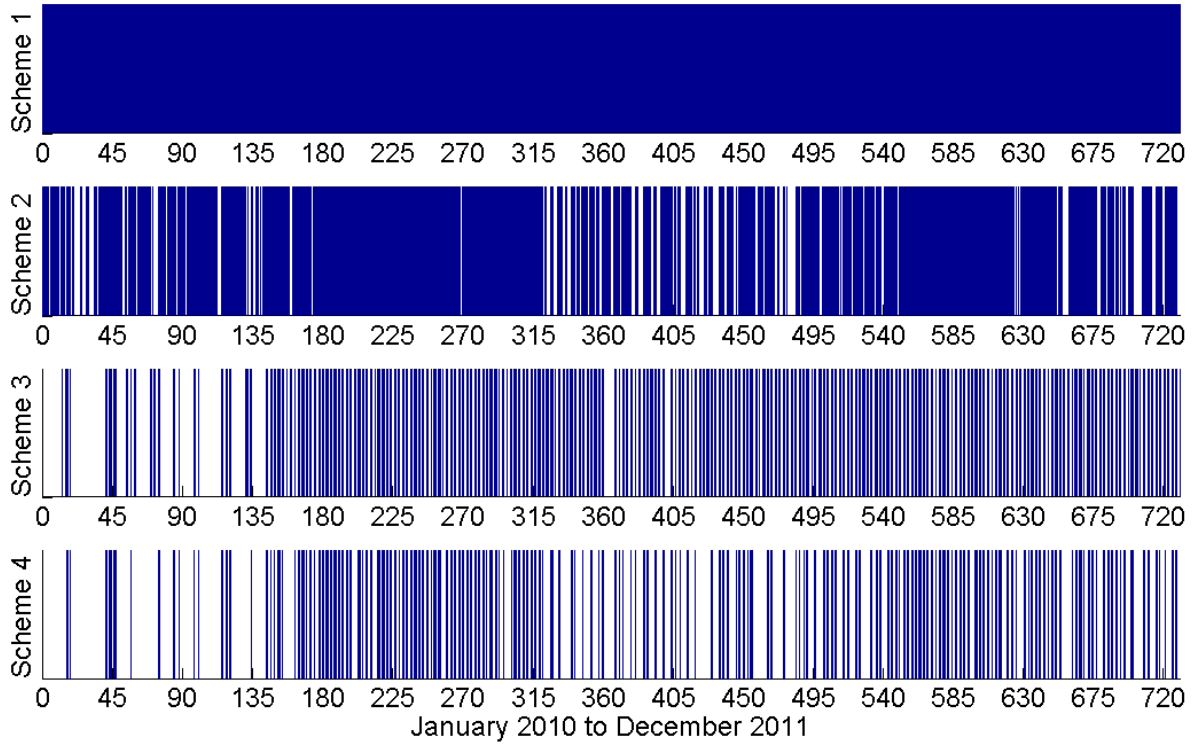
Figure 9. Data availability plots of the four schemes: Scheme 1: SAC-SMA-SM input; Scheme 2: SAC-SMA-SM and MODIS-LST inputs; Scheme 3: SAC-SMA-SM and SMOS-T_bs inputs; Scheme 4: SAC-SMA-SM, MODIS-LST, and SMOS- T_bs inputs. The total available days for the four schemes are 730, 458, 217, and 140 respectively.
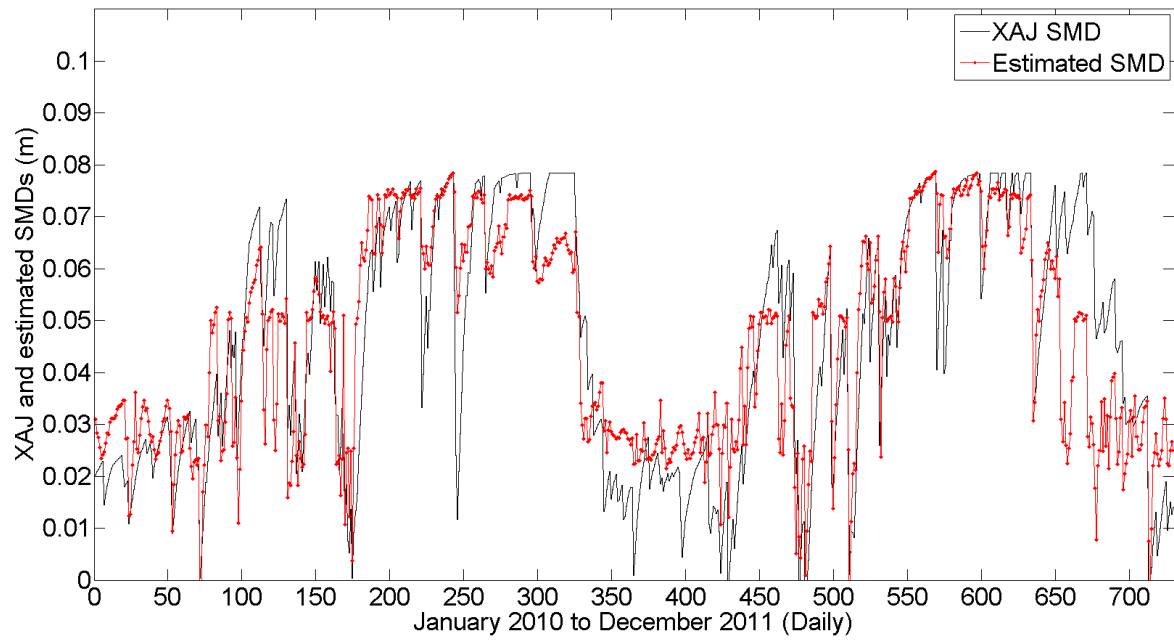
Figure 10. Time series plot of the combined daily hydrological soil moisture state estimations.
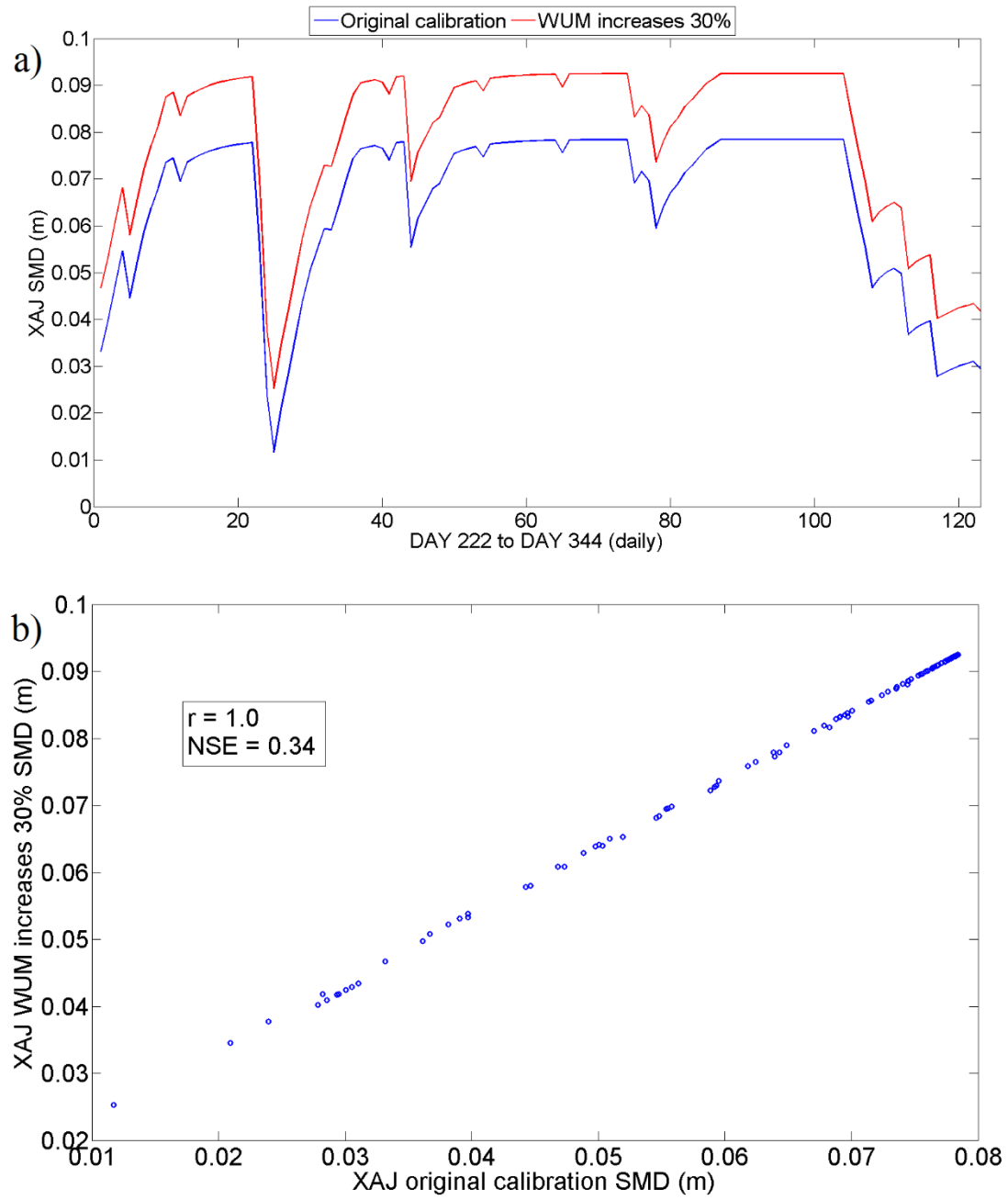
Figure 11. SMD variations from the manipulated XAJ calibration (i.e., the WUM parameter is increased by 30 %) and its original calibration.
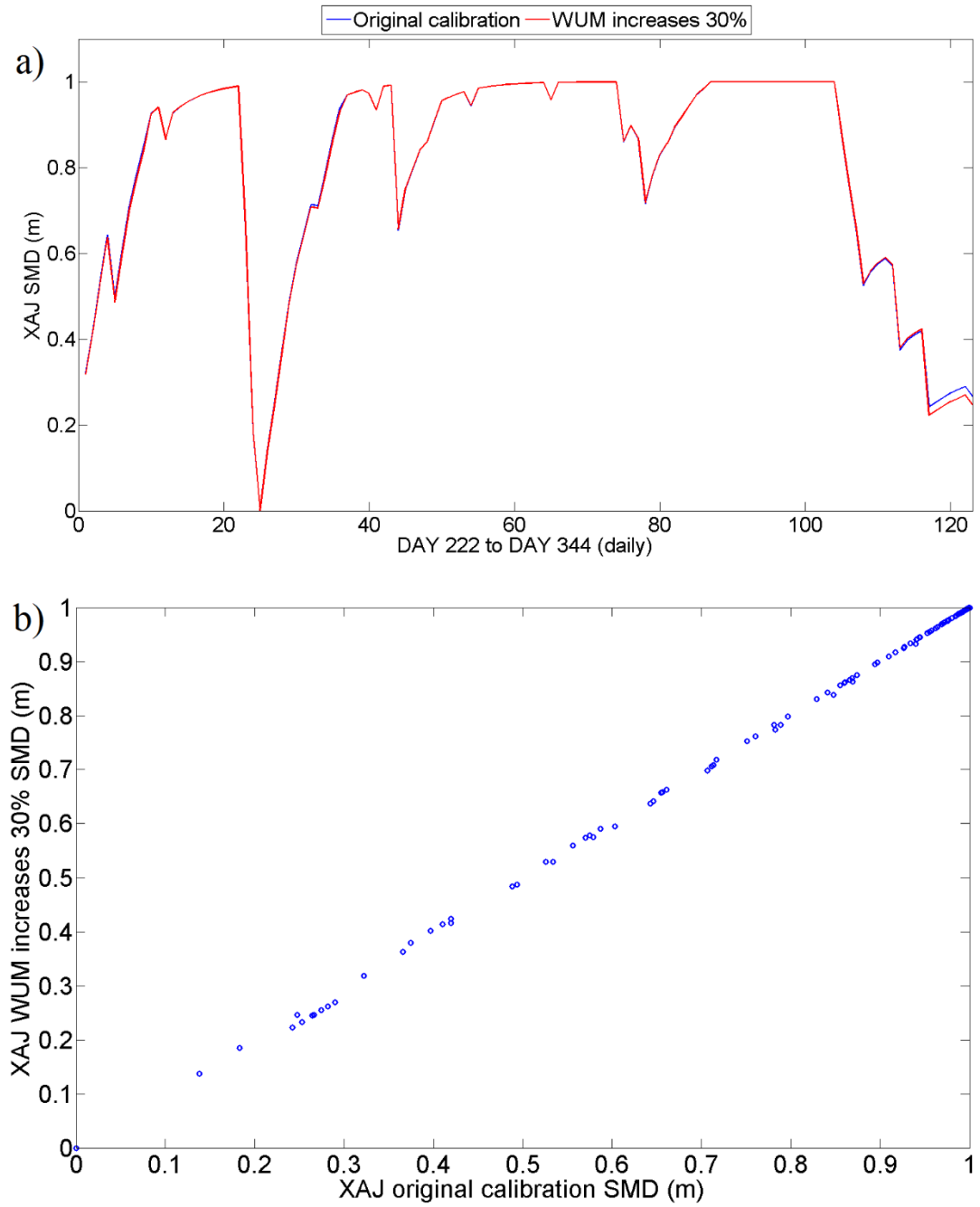
Figure 12. Normalised SMD variations from the manipulated XAJ calibration (i.e., the WUM parameter is increased by 30 %) and its original calibration.

61