Soil Moisture Sensor Network Design for Hydrological Applications

Lu Zhuo[1,2], Qiang Dai[1,3*], Binru Zhao[1], Dawei Han[4]

[1] Key Laboratory of VGE of Ministry of Education, Nanjing Normal University, Nanjing, China

[2] Department of Civil and Structural Engineering, University of Sheffield, Sheffield, UK

[3]Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing, China

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

[*]Correspondence: qd_gis@163.com

Abstract

Soil moisture plays an important role in the partitioning of rainfall into evapotranspiration, infiltration and runoff, hence a vital state variable in hydrological modelling. However, due to the heterogeneity of soil moisture in space most existing in-situ observation networks rarely provide sufficient coverage to capture the catchment-scale soil moisture variations. Clearly, there is a need to develop a systematic approach for soil moisture network design, so that with the minimal number of sensors the catchment spatial soil moisture information could be captured accurately. In this study, a simple and low-data requirement method is proposed. It is based on the Principal Component Analysis (PCA) for the investigation of the network redundancy degree; and *K*-means Cluster Analysis (CA) and a selection of statistical criteria for the determination of the optimal sensor number and placements. Furthermore, the long-term (10-year) 5 km surface soil moisture datasets estimated through the advanced Weather Research and Forecasting (WRF) model are used as the network design inputs. In the case of the Emilia Romagna catchment, the results show the proposed network is very efficient in estimating the catchment-scale surface soil moisture (i.e., with *NSE* and *r* at 0.995 and 0.999, respectively for the areal mean estimation; and 0.973 and 0.990, respectively for the areal standard deviation estimation). To retain 90% variance, a total of 50 sensors in a 22,124 km^2

catchment is needed, which in comparison with the original number of WRF grids (828 grids), the designed network requires significantly fewer sensors. However, refinements and investigations are needed to further improve the design scheme which are also discussed in the paper.

Keywords: Soil moisture network design, Hydrological modelling, Principal Component Analysis (PCA), *K*-means Cluster Analysis (CA), Weather Research and Forecasting (WRF), Optimising, Data mining.

1. Introduction

Soil moisture is at the heart of the Earth system and it plays an important role in the exchanges of water and energy at the land surface (Dorigo et al., 2017;Robock et al., 2000;Crow et al., 2018). In hydrology, soil moisture is the key component for the partitioning of rainfall into evapotranspiration, infiltration and runoff (Vereecken et al., 2008;Brocca et al., 2017;Rajib et al., 2016;Fuamba et al., 2019). In particular, the antecedent soil moisture condition of a catchment is among one of the most important factors for flood triggering (Uber et al., 2018;Zhuo and Han, 2017). For hydrological modelling, soil moisture is a vital state variable. Especially, during real-time flood forecasting, the accurate updating of the soil moisture state variable is a critical step to reduce the accumulation of model errors (i.e., time drift problem) (Lopez et al., 2016;Laiolo et al., 2016;Zwieback et al., 2019). Therefore, the intensive monitoring of catchment-scale soil moisture content would benefit a number of hydrological applications.

In-situ soil moisture sensors (e.g., capacitance probe and Time Domain Reflectometry) can provide point-based soil moisture measurements with relatively high accuracy (after calibration) in comparison with the modelling and the remotely sensed approaches (Albergel et al., 2012). Therefore, they are a crucial source of information for hydrological research (Western et al., 2004;Brocca et al., 2017). However, due to the spatial heterogeneity of soil moisture and the

economic considerations (e.g., installation and maintenance cost), most existing in-situ networks rarely provide sufficient coverage to capture the catchment spatial soil moisture variations (Chaney et al., 2015). There have been enormous works carried out by the USA National Soil Moisture Network (NSMN, 2020), USA state Mesonets and the International Soil Moisture Network (ISMN) (Dorigo et al., 2013) on soil moisture network integration and database setup, however they are based on existing in-situ networks and the majority of which are not purposely designed for catchment scale hydrological studies. In particular, in a number of cases, soil moisture sensors are mainly installed close to the residential plain areas (e.g., due to easy accessibility and maintenance reasons), and there is a lack of sensors installed in the complex topographic areas (Zhuo et al., 2019b).

Therefore, there is a need to develop a systematic approach for the soil moisture network design, so that with the minimal number of sensors the catchment-scale soil moisture information could be captured accurately. Although a number of projects have been carried out in the field of soil moisture network design, for instance through various NASA soil moisture campaigns (SMEX, SMAPVEX, etc.), they are mainly focused on satellite soil moisture evaluations and algorithm improvements, so the in-situ sensors are purposely designed to best match satellite's footprint, with high sensor coverage in small experimental scales. Moreover, most existing soil moisture network studies are based on using in-situ/aircraft datasets at small experimental areas, which can hamper their applications in data-sparse regions. However, to our knowledge, there is a lack of existing literature covering soil moisture network design, and particularly for the catchment-scale hydrological applications (Chaney et al., 2015).

Therefore, to address the aforementioned research gap, the aim of this paper is to propose an efficient soil moisture network design scheme for catchment-scale studies, based on a combination of statistical approaches and globally available modelling datasets. In particular, the Principal Component Analysis (PCA) is adopted to investigate the network redundancy

degree, and *K*-means Cluster Analysis (CA) and a selection of statistical criteria are used to determine the optimal sensor number and placements. Although other statistical methodologies could also be explored (e.g., the Temporal Stability Analysis (Vachaud et al., 1985) and the Empirical Orthogonal Functions (Perry and Niemann, 2007) which have been applied for soil moisture network design by the community), PCA/CA are a simple statistical approach that works efficiently with a large array of datasets and have been successfully explored by Curtis et al. (2019) on classifying soil moisture response units in a catchment. Long-term (10-year) soil moisture datasets estimated through the advanced Numerical Weather Prediction (NWP) Weather Research and Forecasting (WRF) model (Skamarock et al., 2008) are used as the design inputs. WRF model has been applied in a wide range of applications with good performances (Srivastava et al., 2015;Zaitchik et al., 2013;Zhuo et al., 2019a;Stéfanon et al., 2014). Although WRF estimated soil moisture cannot represent the ground truth, they are ideal datasets to provide catchment characteristics, such as land cover, soil properties, topographies, which are the main drivers of local soil moisture heterogeneity (Friesen et al., 2008). Therefore, such globally available datasets together with the proposed statistical approaches would provide useful insights for the soil moisture network design research (i.e., to minimise the redundancy of information, and improve accuracy), in particular, for those currently ungauged catchments. In this study, the proposed method is implemented in the Emilia Romagna region, northern Italy as a case study due to its high-exposure of flood events.

The paper is organised as: the study area is introduced in Section 2; soil moisture network design methodologies are described in Section 3; the results are presented in Section 4; and discussions and conclusions are included in Section 5.

2. Study Area

In this study, the Emilia Romagna region (latitude 43º50ˈN–45º00ˈN; longitude 9°20ˈE–12°40ˈE) is selected for the case study which is in Northern Italy (Figure 1). The region's total coverage is approximately 22,124 km^2. It is surrounded by the Apennines to the south and the Adriatic Sea to the east, with over half of the area as a plain agricultural zone (12,000 km^2). The climate condition is highly varied in the region which is largely influenced by the mountains and the sea, with subcontinental in the Po river plain and surrounding hilly areas, and cool temperate in the mountain range (Nistor, 2016). It has distinct wet and dry seasons (i.e. dry season between May and October, and wet season between November and April) (Zhuo et al., 2019b). Based on the ESA Climate Change Initiative land cover map (Bontemps et al., 2013), the region is mainly covered by Herbaceous (37%), followed by Tree (22%), and Cropland (21%). The majority of the area is on the quaternary alluvial deposits, which are characterised by a high degree of heterogeneity (Pistocchi et al., 2015). The annual temperature ranges from 8.2 to 19.3ºC; and the annual mean precipitation is between 520 and 820 mm (Pistocchi et al., 2015).

For the soil moisture network in the region, currently, there is a total of 19 soil moisture sensors installed (all located in the plain area); however only one of them can provide long-term continuous soil moisture monitoring datasets. The network is managed by the Regional Agency for Environmental Protection Emilia Romagna Region. Through further investigations, it has been found, a number of the sensors have actually never provided proper soil moisture measurements since the installation. Only one soil moisture sensor at the plain area is clearly not sufficient for any catchment-scale applications. Therefore, it is hoped the proposed soil moisture network design scheme could provide some useful guidance to the local authority on an improved network in the future (i.e., a minimum number of sensors for reduced installation and maintenance cost, but at the right locations).

3. Methodologies

3.1 WRF Model

The WRF model is a next-generation, non-hydrostatic mesoscale NWP system designed for both atmospheric research and operational forecasting applications (Skamarock et al., 2005). The model is capable of modelling a wide range of meteorological applications varying from tens of metres to thousands of kilometres (NCAR, 2018). Apart from the WRF's aforementioned advantage on including the catchment characteristics for the soil moisture estimations, it also has other merits that make it an ideal tool for providing the distributed soil moisture information for the network design. For instance, WRF model's spatial and temporal resolutions can be changed depending on the input datasets to fit various application requirements, and a number of globally-available data products can be selected to provide the necessary boundary and initial conditions for running the model. Therefore, WRF is able to provide valuable information for this study. Here WRF version 3.8 is used.

3.1.1 Model Parameterization

Apart from the atmospheric forcing, parameterization is also required to drive the WRF model. In particular, the microphysics scheme is important in simulating accurate rainfall information which in turn is significant for estimating the accurate soil moisture fluctuations. WRF V3.8 supports 23 microphysics options ranging from simple to more sophisticated mixed-phase physical options. In this study, the WRF Single-Moment 6-class scheme is adopted which considers ice, snow and graupel processes and is suitable for high-resolution applications (Zaidi and Gisen, 2018). The physical options used in the WRF setup are Dudhia shortwave radiation (Dudhia, 1989) and Rapid Radiative Transfer Model (RRTM) longwave radiation (Mlawer et al., 1997). Cumulus parameterization is based on the Kain-Fritsch scheme (Kain, 2004) which is capable of representing sub-grid scale features of the updraft and rain processes, and such a feature is useful for real-time modelling (Gilliland and Rowe, 2007). The surface layer

parameterization is based on the Revised fifth-generation Pennsylvania State University–National Center for Atmospheric Research Mesoscale Model (MM5) Monin-Obukhov scheme (Jiménez et al., 2012). The planetary boundary layer is calculated based on the Yonsei University scheme (Hong et al., 2006jimenez). In WRF, its land surface model plays a vital role in the integration of information generated through the surface layer scheme, the radiative forcing from the radiation scheme, the precipitation forcing from the microphysics and convective schemes, and the land surface conditions to simulate the water and energy fluxes (Ek et al., 2003). In this study, the Noah Multiparameterization (Noah-MP) is chosen, because it has shown more accurate soil moisture estimation performance than the other two main schemes (Noah and CLM4) in other studies (Cai et al., 2014;Zhuo et al., 2019a). Table 1 shows the selected WRF parameterization schemes. The static inputs (i.e., land use and soil texture) are chosen in the WRF pre-processing package. Here, the land use categorisation is interpolated from the MODIS 21-category data classified by the International Geosphere Biosphere Programme (IGBP). The soil texture data are based on the Food and Agriculture Organization of the United Nations Global 5-minutes grid cell soil database.

3.1.2 Model Setup

The WRF model is centred over the Emilia Romagna Region, and integrates three nested domains (D1, D2, D3), with the horizontal spacing of 45 km x 45 km (outer domain, D1), 15 km x 15 km (inner domain, D2), and 5 km x 5 km (innermost domain, D3). In this study, the innermost domain D3 is used (88 x 52 grids [west-east and south-north, respectively]), with a two-way nesting scheme considered letting the information from the child domain to be fed back to the parent domain. To drive the WRF model, the European Centre for Medium-Range Weather Forecasts (ECMWF) reanalysis (ERA-Interim) is adopted to provide the study region's boundary and initial conditions. ERA-Interim is a global atmospheric reanalysis that is available from 1979 to 2019 (ERA-5 as a recent update to ERA-Interim may also be used).

The spatial resolution of the datasets is approximately 80 km on 60 levels in the vertical from the surface up to 0.1 hPa. It contains 6-hourly gridded estimates of three-dimensional meteorological variables, and 3-hourly estimates of a large number of surface parameters and other two-dimensional fields. Please see (Berrisford et al., 2011) for a detailed documentation of the ERA-Interim.

After the initialization, the model needs to be spun-up to derive a physical valid state (e.g., equilibrium state) (Cai et al., 2014;Cai, 2015). In this study, WRF is spun-up by running through the whole year of 2005. After the spin-up, the WRF model is run in daily timestep from January 1, 2006, to December 31, 2015, using the ERA-Interim datasets. The modelled WRF grids within the Emilia Romagna catchment (total of 828 grids) are shown in Figure 2 as black dots, with the elevation map also illustrated in the background. For the exploration purpose, this study uses the WRF surface soil moisture at 0-10 cm for the network design. This is because the surface soil moisture changes more frequently in comparison with the root-zone soil moisture. And in theory, the root-zone soil moisture should follow the general trend of the surface soil moisture (in a delayed mode). In our future study, the WRF root zone soil moisture will also be explored.

3.2 Soil Moisture Network Design

For the soil moisture network design, three main problems need to be tackled. First is how redundant the network is, second is how many soil moisture sensors are needed within a catchment, and finally where are the best locations to place them. To solve the first problem, the PCA is used to investigate the redundancy degree of the network. To solve the latter two problems, the *K*-means CA is adopted. It should be noted that the information used for the PCA/CA is based on the soil moisture temporal variations (e.g., the 10-year time series data), so that areas following similar soil moisture variations can be grouped together, and location

information is not used here. However, due to the influence of local characteristics, the resultant clusters should more or less reflect the geographical feature.

3.2.1. Principal Component Analysis (PCA) for Network Redundancy Analysis

When soil moisture data are collected from p soil moisture sensors, these data are often correlated. This correlation reflects the complexity of the catchment and indicates that some of the information collected from one sensor is also contained in the remaining p-1 sensors (Gangopadhyay et al., 2001). The role of the PCA is to examine the redundancy of the WRF 5 km gridded soil moisture outputs (Dai et al., 2017). PCA is a statistical procedure for multivariance feature extraction. It adopts an orthogonal transformation to convert a set of possibly correlated observations into a set of linearly uncorrelated variables called principal components. This transformation is defined in such a way that the first principal component has the largest possible variance, and each succeeding component in order has the highest variance possible under the constraint that it is orthogonal to the preceding components (Wold et al., 1987).

In this study, we have p WRF soil moisture grids with N observations (the time series of the data, i.e., 10-year daily datasets). The covariance matrix p x p can be calculated which is denoted as X, and the eigenvectors and the eigenvalues of the matrix can also be determined, correspondingly. Since eigenvectors of X are orthogonal, the p eigenvectors are used to construct the principal components, which can be represented as:

$$\text{eigenvector} = (eig_1 \ eig_2 \ eig_3 \ \dots \ eig_p) \quad (1)$$

with such a relationship, the original datasets can be transformed in terms of eigenvectors into a new dataset Z. Z is shown as the following:

$$Z_i = X_1 eig_{i,1} + X_2 eig_{i,2} + \dots + X_p eig_{i,p} \ , \quad i = 1, \dots, p \quad (2)$$

where Z_i is the new dataset, X_i is the original dataset. The variance of each of the component is the eigenvalue. The eigenvector with the highest eigenvalue is the principal component of the dataset. Since the optimal number of principal components is dependent on the amount of original variance the network should retain, the examination of the network redundancy is implemented based on the desired rate of variance contribution, and the number of principal components can thus be calculated correspondingly.

3.2.2. *K*-means Cluster Analysis (CA) for Sensor Number and Placements Determination

After exploring the redundancy level of the network, it is necessary to determine how many WRF grids to select so that the maximum level of information can be retained. Similar to the relationship between the number of principal components and the variance contribution rates, the appropriate number of grids are also dependent on the amount of original variance the network would like to retain. Since the number of components from the PCA do not directly represent the physical number of grids, we propose to use the elbow method to find the corresponding number of grids. The elbow method is based on *K*-means clustering and looks at the variance contribution rate as a function of the number of grids. Generally, the required number of grids increases when the variance contribution rate increases. However, the growth rate is not constant and changes significantly at a critical point (threshold), which is used in this study as the desired rate for the soil moisture network design. If for a specific desired variance, the determined number of grids is significantly less than the total number of the WRF soil moisture grids, then it can be concluded that the network is heavily redundant, and even by removing a large number of grids, the remaining can still provide sufficient soil moisture information for the entire catchment; and vice versa. In this paper, the variance contribution rate of 70%~99% is tested.

K-means approach is a typical distance-based clustering method which uses the distance as the indicator for similarity among objects (i.e., the smaller the distance, the higher the similarity of two objects) (Kodinariya and Makwana, 2013). In this study, the Euclidean distance is adopted as the distance measurement. It is a simple and widely used way of calculating the distances between objects in a multidimensional space (Danielsson, 1980). The centroid of each cluster is the point which the sum of Euclidean distances from all objects in that cluster is minimized. It is an iterative approach repeated for all of the clusters.

After deciding the number of soil moisture grids from the elbow method and setting up the optimal clusters, we propose two ways to find the most suitable grid for the sensor placements. One way is by finding the grid which gives the median averaged soil moisture (i.e., averaged over the whole study period) in each of the cluster (denoted as CA-Med), and another is through identifying the maximum averaged soil moisture in each of the cluster (denoted as CA-Max) (Dai et al., 2017). The CA-Max is focused on extreme soil moisture conditions, whilst the CA-Med is on the median condition. Since they provide results in two aspects, it is useful to explore both in this study. As a result, for each cluster, there is one optimal grid, and grouped with the other optimal grids found in other clusters, the ideal placements for the soil moisture sensors are identified. The group of the selected grids is considered to be the optimal combination of locations that can provide the desired variance of the original WRF soil moisture measurements over the whole catchment.

3.3 Network Evaluation

Since there is no existing optimal in-situ soil moisture network that can be used as a reference for the evaluation, it is challenging to assess the designed network performance based on a comparison study. However, the designed network should be efficient enough to represent the maximum amount of information with the minimum number of sensors within a catchment. In other words, the designed network should retain the main catchment-scale soil moisture

information of the original WRF's full outputs, which is particularly important for the hydrological modelling. To assess the network in such an aspect, the soil moisture information contained by the designed and the original network are compared. Two statistical indicators are used for the purpose, namely the Pearson correlation coefficient and the Nash–Sutcliffe coefficient.

The Pearson correlation coefficient (r) is a statistical measure of the linear correlation between two sets of datasets, which in this study can estimate the systematic deviation between the designed (S_d) and the original (S_o) catchment-scale soil moisture variations, and it is calculated by the following equation:

$$r_{S_o,S_d} = \frac{E[S_d S_O] - E[S_d]E[S_O]}{\sqrt{(E[{S_d}^2] - E[S_d]^2) \times (E[{S_O}^2] - E[S_O]^2)}} \quad (3)$$

where E is the mean value of the corresponding vector. In this study, the optimal performance is achieved when r_{S_o,S_d} equals to 1

Nash-Sutcliffe Efficiency (*NSE*) (Nash and Sutcliffe, 1970) is used widely in hydrology to evaluate the prediction accuracy in hydrological modelling, which can be obtained by:

$$NSE = 1 - \frac{\sum (s_o^t - s_d^t)^2}{\sum (s_o^t - E[s_o])^2} \quad (4)$$

where t is the time-step of the dataset. The *NSE* ranges [1,-∞). The closer *NSE* is to 1, the more accurate the designed network is.

4. Results

4.1. Soil Moisture Network Redundancy Analysis

Within the study area of 22,124 km^2, there is a total number of 828 WRF soil moisture grids. With such a dense dataset, there should exist information redundancy. To explore this, a cross-

correlation (r) matrix for all of the grids over the whole study period is plotted in Figure 3. It can be seen that the majority part of the matrix is in blue-tone, which means most of the grids (85%) are correlated ($r > 0.5$) with most of the others (as shown in Table 2). In addition, over half of the grids (52%) have high correlation ($r>0.8$) with the rest of the grids; and even 15% of the grids can achieve very high correlation ($r>0.9$). However, it is clear from the map some grids (e.g., grid number 396-398, 523-529) are less strongly correlated with the others (red-tone, with low correlation <0.3 observed), which means more soil moisture sensors might need to be installed in those locations. A further exploration of cross-correlation performance using box plots is shown in Figure 4b). The locations of the selected grids (as in Figure 4b) are marked in Figure 4a) with red circles. It can be seen the nine grids are distributed evenly within the catchment in order to represent a spectrum of catchment features (e.g., different land covers, elevations, soil types etc.). From the box plot, it can be seen for a specific grid, the cross-correlation can range from as low as below 0.1 to as high as almost 1. The large range is particularly obvious for Grid 500, which is located at the plain zone near the east boundary of the catchment and is close to the Valli di Comacchio lagoon. The closeness to the waterbody could mean its soil moisture is dominated more by the shallow water table at that location which makes the soil moisture relatively insensitive to the weather, in comparison with grids located further away. For Grid 100, its correlation with the rest of the grids in the catchment is relatively low, with 75% percentile of the cross-correlations less than 0.6. The potential reason could be because it is located in the southern mountainous zone, with high-density of tree coverage and complex topographic conditions, its soil moisture changes more differently than the other grids. A similar condition is observed for Grid 1 which is also located in a hilly zone in the southern boundary of the catchment (i.e., lower correlation as shown in the boxplots). Such a phenomenon is not unexpected and could mean more sensors are needed in those complex zones for better soil moisture monitoring purpose. However, for Grids like 300, and

600 (and the surrounding areas), since the majority of their correlations are high and they are located in plain areas with no water boundary nearby, they could be arranged with a smaller number of soil moisture sensors.

4.2. PCA Analysis and Sensor Number

In summary, through the cross-correlation exploration, many parts of the WRF soil moisture dataset are significantly redundant. To systematically investigate the redundancy degree of the network, the PCA approach is applied. Figure 5a) shows the PCA results to provide useful guidance on the acceptable loss of information. It is clear to see the first principal component carries close to 80% of the total variance, with the second component bringing this to nearly 90%. This result again indicates the high redundancy exists in the dataset, and just one component can contain almost 80% of the total soil moisture information. To better understand the relationship between the principal component numbers, the variance contribution rate, as well as the corresponding required grids number (through elbow method), a set of variance contribution rates from 70% to 97.5% is used as the representatives. The required number of components and the grids are listed accordingly in Table 3. It can be seen only one component with 6 grids is sufficient to retain 70% of the soil moisture information. Even when the variance is set at 80%, only two components are needed to meet the requirement, and the corresponding number of soil moisture girds is 11 (1.3% percent of the total grids). To satisfy 90% variance, three components are needed, and although the total number of grids is increased to 50, it is still significantly less than the WRF's full inputs. The detailed numbers further indicate the relatively high level of redundancy in the WRF's original dataset.

The trend can also be observed through the elbow curve which is illustrated in Figure 5b). It presents the relationship between the variance and the number of grids. It can be seen to meet the increment of variance, the required number of grids also increases. But the growth rate is

the most significant when the variance is smaller than 70% and then slows down gradually after that. When the variance meets 95%, the rate is further weakened. Based on the curve, it is suggested the desired variance (i.e., trade-off point) between 80% and 95%. The required number of soil moisture grids for 80%, 85%, 90%, and 95% is 11, 21, 50, and 184 respectively. It is clear, in order to achieve the 95% variance, a significantly greater number of additional grids are required, that is 268% more than for the 90% variance case. Therefore, for further improvement of variance from 90% to 95%, the economic cost for the additional number of sensors might not be as valuable as for the 85% to 90% case (138% additional sensors are required for the enhancement).

4.3. Soil Moisture Sensor Location Design

Once the degree of redundancy for the full WRF soil moisture network is established, the next step is to determine the optimal locations for sensor placements. Here, CA-Max and CA-Med are used. The designed networks for CA-Max and CA-Med are illustrated in Figure 6 and 7, respectively. The indicated locations in the figures provide guidance on the preferential areas for the soil moisture sensor placements. Each of the methods gives a different set of sensor locations, for instance, the selected optimal soil moisture grids from the CA-Max method tend to be located at the catchment boundary, and the situation is particularly obvious for the low variance cases (i.e., 70% - 80%). For example, when the variance is set at 70%, the selected optimal locations from the CA-Max is mostly distributed near the catchment's southern boundary, while from the CA-Med, it is more homogeneously distributed (i.e., one at the southern boundary, one at the north, two at the north-western part, and two at the north-eastern part). This is because CA-Max selects the maximum averaged soil moisture of a cluster. In the case study area, since the southern boundary of the catchment is mainly covered by dense tree which generally has higher soil moisture contents than the rest of the catchment, the selected

363 locations tend to distribute near the southern boundary. For the CA-Med, as it selects the
364 median averaged soil moisture of a cluster, the resultant locations are more evenly distributed.

365 When the variance is increased, for instance at 90%, the difference between the two CA
366 methods becomes less distinctive. Despite this, it can still be seen for the CA-Max, there is less
367 coverage of sensors at the western and the eastern parts of the catchment, with most of the
368 sensors located at the mid-region. However, for the same variance, the sensor distribution from
369 the CA-Med looks more evenly distributed visually. Nevertheless, when the variance reaches
370 as high as 97.5%, the difference from the two methods becomes rather small, as 367 sensors
371 are located covering most parts of the catchment in both cases.

372 **4.4. Soil Moisture Network Evaluation**

373 The evaluation of the designed network is challenging, as there are no standard assessment
374 criteria available to guide on what kind of network is the most appropriate for a given study
375 area. In essence, the designed network should be efficient, which means the network should
376 contain the maximum amount of information with a minimal number of sensors. In this study
377 since we focus on the soil moisture's hydrological applications (catchment-scale), to evaluate
378 the efficiency of the proposed schemes, the catchment-scale soil moisture data derived by the
379 designed networks are compared with the WRF's full inputs (828 grids). Both the areal spatial
380 mean and standard deviation are calculated. The Pearson correlation coefficient and the Nash–
381 Sutcliffe coefficient are used to quantify the relationships between the two soil moisture
382 datasets. The results for both the CA-Med and the CA-Max are compared in Figure 8. Based
383 on the areal mean soil moisture (Figure 8 a) and c)), it is clear to see the CA-Med outperforms
384 the CA-Max for the majority of the variance cases (both *NSE* and *r*), except for the *NSE* results
385 when the variance is over 90%. Moreover, for the *NSE* results, a decline of the performance
386 can be observed clearly after it passes the 90% variance point, which illustrates that an

increment of sensor number does not necessarily mean a arise of the performance. For the standard deviation, the disparity between the two methods is smaller. When the variance is below 80%, the growth trend for the CA-Med case is not clear, as it firstly drops at the 75% point and then climbs up again when the variance increases. Whereas for the CA-Max case, there is a clear upward trend. Similar to Figure 8 a), it is interesting to see for the areal standard deviation in Figure 8 b) and d), the *NSE* and *r* also start to drop after reaching around 90%. The evaluation results are summarised in Table 4 for numerical comparison. Since CA-Med surpasses CA-Max for most of the cases, it is chosen for the network design. In the aspect of the desired variance, because as discussed earlier, when the variance climbs over 90%, the performance instead drops. Therefore 90% variance is suitable to be used for the network design in this case.

The time series plots of the areal soil moisture mean and standard deviation are shown in Figure 9. Generally, the designed network can estimate the catchment's mean soil moisture very well, as it follows the variation of the WRF's full input dataset closely (*NSE* = 0.995 and *r* = 0.999). For the standard deviation, the general trend from both datasets shows a higher spatial variation of soil moisture over the dry season and lower variation during the wet season. The spatial variation is averaged around 0.04 m^3/m^3 throughout the whole study period. However, there are some disparities between the two datasets, in particular, during the wet season (bottom peaks in the STD plot), the designed network at several occasions overestimates the spatial soil moisture variation, and during the dry season (top peaks in the STD plot), it underestimates instead. Nevertheless, the differences are small and the correlation between the two datasets is high, with *NSE* = 0.973 and *r* = 0.990 obtained. In conclusion, the designed network can maintain the dominated information of the WRF's full-grid input well.

The sensor displacements for the designed and the existing (in-situ) networks are illustrated in Figure 10. In comparison with the distribution of the proposed network, the existing network

412 is clearly biased, with all of the sensors located in the mid-plain zone only. Such distribution
413 (i.e., no sensors located at the southern mountainous (highly-vegetated) region) can have
414 adverse impacts on the accuracy of the areal mean soil moisture estimation. However, we can
415 see some of the existing sensors are located near some of the designed sensors, which could be
416 kept if located within the same cluster. But a lot more sensors are indeed required in the hilly
417 zone, where currently no sensors are installed. The existing stations could be initially installed
418 for irrigation purpose, which are hence mainly located in the plain area. Scatterplots of the areal
419 mean soil moisture calculated from the designed and the existing networks are also presented
420 in Figure 11. The performance difference between the two networks is clear to observe. For
421 the proposed network, the points are located close to the identical line, whereas for the existing
422 network, due to the inappropriate sensor distributions over the catchment, the points are more
423 dispersive (*NSE* = 0.889). The performance of the existing network (i.e., using WRF grid data
424 from the existing locations) in comparison with the proposed networks is worse, in particular,
425 its *NSE* is lower than the 70% CA-Med case in the designed network (i.e., 0.949). For the
426 existing network, without putting sensors in the highly vegetated region, the network clearly
427 underestimates soil moisture variations during the dry season (i.e., for the cases when the soil
428 moisture is less than 0.25 m^3/m^3)

429 **5. Discussions and Conclusions**

430 With the low-cost soil moisture sensors becoming more and more available and modern
431 communication technology (i.e., Internet of Things), it is expected more in-situ soil moisture
432 sensors will be installed in the future. Although there is a wide range of soil moisture networks
433 around the world (e.g., USA NSMN, ISMN, USA state Mesonets), majority of them are not
434 purposely designed for catchment scale hydrological studies. Moreover, to our knowledge most
435 existing soil moisture network studies are based on using in-situ/aircraft datasets at small
436 experimental areas, which can hamper their applications in data sparse regions. In this paper, a

low-data requirement scheme (only WRF simulated soil moisture information is required, which can be generated globally) together with simple statistical analysis (PCA/CA) is proposed to overcome the aforementioned shortcomings. Through a series of evaluations of the developed network, it can be concluded that the method can provide efficient catchment-scale soil moisture estimations (i.e., high accuracy of the areal mean and standard deviation soil moisture estimations). To retain 90% variance, a total of 50 sensors in a 22,124 km^2 catchment is needed. In comparison with the original number of WRF's grids (828 grids), the proposed network requires significantly smaller number of sensors. Furthermore, in comparison with the existing soil moisture network in the Emilia Romagna region, the proposed network has sensors more evenly distributed, covering most representative parts of the catchment (e.g., both plain and mountainous regions), and can obtain more accurate catchment-scale soil moisture estimation. However, there are several points need to be discussed as follows.

The first point is about the uncertainty of the WRF's soil moisture estimations, which could influence the accuracy of the network design. It is acknowledged that the reliability of the designed network is influenced by the performance of the WRF model. To evaluate the WRF results and test whether the proposed network can produce the catchment-scale soil moisture well, a long-term densely covered soil moisture network will be required. Setting up such a network is challenging and difficult to realise due to the high installation and maintenance cost. In this study, a long-term WRF soil moisture estimation with 1-year spin-up time is used which could to some extent produce a more stable result. But since "all models are wrong" (by George E. P. Box), an uncertainty model (Zhuo et al., 2016) could be proposed to be integrated with the network design scheme. For example, we can generate a large number of probable "true soil moisture" datasets based on the proposed uncertainty model so that a set of possible soil moisture networks can be produced. As a result, the designed network will be expressed in a probabilistic form instead of a determinate form. In addition, a decision-making scheme

462 considering different conditions (e.g., accessibility, installation and maintenance cost) under
463 the uncertainty can be developed to select the most suitable soil moisture network. The
464 uncertainty influence of the WRF soil moisture on the network design will be investigated in
465 future studies.

466 Second, the case study is based on the daily soil moisture inputs for the hydrological
467 applications. With different research needs (meteorology, climatology, hydrology, water
468 resources, geology, etc.), various temporal-scale of soil moisture data might be required, for
469 example, climate change study requires soil moisture data in decades or hundreds of years
470 which often needs annual-scale measurements; drought assessment requires monthly to
471 seasonal datasets; while for hydrometeorological prediction applications, hourly datasets might
472 be needed. For the network design, the input data's temporal scale (daily, weekly, monthly,
473 yearly) can influence the final network design, therefore it is worth investigating in future
474 studies about the temporal-scale effect on the network design.

475 Third, for a complex catchment like Emilia Romagna, other uncertainty sources apart from the
476 WRF model can also affect the performance of the designed network; for instance, the study
477 area has varied climate conditions (a mixture of subcontinental and cool temperate) and distinct
478 seasonal changes (wet/dry seasons). Therefore separating/combining networks under different
479 catchment conditions could result in an improved soil moisture network design. Furthermore,
480 the poor accessibility to sensors is another challenging point that can hamper the performance
481 of the designed network in real life. To overcome the accessibility issue, advanced soil moisture
482 sensors (e.g., Cosmic-ray soil moisture sensor (Zreda et al., 2012)) with low maintenance
483 requirement, could provide good alternative for long-term deployment in complex terrain.
484 Moreover, the accessibility factor could also be considered for the network design (e.g., can be
485 considered during the CA for the sensor placements).

Fourth, the proposed method assumes that a soil moisture station placed inside a 5-km grid cell can perfectly represent the mean soil moisture condition for that grid cell. However, in reality it is not the case. As a result, the scale mismatch between the footprint of an in situ point-based soil moisture station and the 5-km data set used here would be expected to degrade the performance of the resulting network (Crow et al., 2012). Advanced soil moisture sensing technology as aforementioned such as the Global Navigation Satellite Systems (GNSS) and the Cosmic-ray could provide alternative solutions over point-based sensors to reduce such impacts. In particular, COSMOSUK (Evans et al., 2016) network is moving towards integration with operational weather forecasts, and Cosmic-ray is suitable to be used in complex terrain and might be a good option to be used for national network as compared with point-based in-situ sensors.

Fifth, regarding the temporal variation factor, as has been mentioned earlier that the information we used for the PCA/CA is based on the soil moisture temporal variations, so that areas following similar soil moisture temporal variations can be identified. Although location information is not used for the PCA/CA analysis, due to the influence of local characteristics, the resultant clusters should more or less reflect the geographical feature. The resultant clusters are shown in Figure 12. It can be seen most of the clusters are geographically connected. Although *k*-means has issues in dealing with nonconvex clusters and geographically there might exist nonconvex shaped clusters, as demonstrated in this paper *k*-means indeed is very useful for the soil moisture network design (the time series datasets are used instead of the location information).

Since the forcing data for the WRF model is globally covered, the proposed scheme can largely benefit ungauged catchments. On the other hand, in places where dense soil moisture networks are already available, the proposed scheme could also help in minimizing the cost by reducing the number of sensors. Another advantage of the method is that the number of soil moisture

sensors can be changed based on different variances to meet various requirements. Through selecting different variance levels, the redundancy of the WRF's full-input network can be assessed, and the corresponding optimal sensor number can be determined. However, the proposed scheme is still in its infancy with a lot of refinements and further explorations needed, therefore it is hoped this paper will stimulate more studies by the community in tackling the soil moisture network design problem.

Data availability

The ERA-Interim data for the WRF modelling can be downloaded from the ECMWF website https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era-interim.

Author contribution

Lu Zhuo carried out the modelling of WRF, build up the soil moisture networks, and prepared the manuscript with contributions from all co-authors. Binru Zhao processed the WRF soil moisture outputs and carried out quality checks. Qiang Dai and Dawei Han provided guidance on the paper's main research direction and are the funding holders of this project in China and UK, respectively.

Competing interests

The authors declare that they have no conflict of interest.

Acknowledgement

This research is supported by the National Natural Science Foundation of China (NSFC, grant no. 41871299), and Resilient Economy and Society by Integrated SysTems modelling (RESIST), Newton Fund via Natural Environment Research Council (NERC) and Economic and Social Research Council (ESRC) (NE/N012143/1).

References

534 Albergel, C., De Rosnay, P., Gruhier, C., Muñoz-Sabater, J., Hasenauer, S., Isaksen, L., Kerr,
535 Y., and Wagner, W.: Evaluation of remotely sensed and modelled soil moisture products using
536 global ground-based in situ observations, Remote Sens. Environ., 118, 215-226, 2012.

537 Berrisford, P., Dee, D., Poli, P., Brugge, R., Fielding, K., Fuentes, M., Kallberg, P., Kobayashi,
538 S., Uppala, S., and Simmons, A.: The ERA-Interim archive, version 2.0,
539 https://www.ecmwf.int/node/8174, 2011.

540 Bontemps, S., Defourny, P., Radoux, J., Van Bogaert, E., Lamarche, C., Achard, F., Mayaux,
541 P., Boettcher, M., Brockmann, C., and Kirches, G.: Consistent global land cover maps for
542 climate modelling communities: Current achievements of the ESA's land cover CCI,
543 Proceedings of the ESA Living Planet Symposium, Edimburgh, 2013, 9-13,

544 Brocca, L., Ciabatta, L., Massari, C., Camici, S., and Tarpanelli, A.: Soil moisture for
545 hydrological applications: open questions and new opportunities, Water, 9, 140, 2017.

546 Cai, X., Yang, Z. L., Xia, Y., Huang, M., Wei, H., Leung, L. R., and Ek, M. B.: Assessment of
547 simulated water balance from Noah, Noah-MP, CLM, and VIC over CONUS using the
548 NLDAS test bed, J. Geophys. Res. Atmos., 119, 13,751-713,770, 2014.

549 Cai, X.: Hydrological assessment and biogeochemical advancement of the Noah-MP land
550 surface model, Doctor of Philosophy, Geological Sciences, The University of Texas at Austin,
551 164 pp., 2015.

552 Chaney, N. W., Roundy, J. K., Herrera-Estrada, J. E., and Wood, E. F.: High-resolution
553 modeling of the spatial heterogeneity of soil moisture: Applications in network design, Water
554 Resour. Res., 51, 619-638, 2015.

555 Chen, F., and Dudhia, J.: Coupling an advanced land surface-hydrology model with the Penn
556 State-NCAR MM5 modeling system. Part I: Model implementation and sensitivity, Monthly
557 Weather Review, 129, 569-585, 2001.

558 Crow, W., Chen, F., Reichle, R., Xia, Y., and Liu, Q.: Exploiting soil moisture, precipitation,
559 and streamflow observations to evaluate soil moisture/runoff coupling in land surface models,
560 Geophys. Res. Lett., 45, 4869-4878, 2018.

561 Crow, W. T., Berg, A. A., Cosh, M. H., Loew, A., Mohanty, B. P., Panciera, R., de Rosnay, P.,
562 Ryu, D., and Walker, J.: Upscaling sparse ground-based soil moisture observations for the
563 validation of coarse-resolution satellite soil moisture products, Rev. of Geophys., 50, 2012.

564 Curtis, J. A., Flint, L. E., and Stern, M. A.: A Multi-Scale Soil Moisture Monitoring Strategy
565 for California: Design and Validation, J AM Water Resour. AS, 55, 740-758, 2019.

566 Dai, Q., Bray, M., Zhuo, L., Islam, T., and Han, D.: A scheme for rain gauge network design
567 based on remotely sensed rainfall measurements, J. Hydrometeorol., 18, 363-379, 2017.

568 Danielsson, P.-E.: Euclidean distance mapping, Computer Graphics and Image Processing, 14,
569 227-248, 1980.

570 Dorigo, W., Xaver, A., Vreugdenhil, M., Gruber, A., Hegyiova, A., Sanchis-Dufau, A.,
571 Zamojski, D., Cordes, C., Wagner, W., and Drusch, M.: Global automated quality control of in
572 situ soil moisture data from the International Soil Moisture Network, Vadose Zone J., 12, 2013.

573 Dorigo, W., Wagner, W., Albergel, C., Albrecht, F., Balsamo, G., Brocca, L., Chung, D., Ertl,
574 M., Forkel, M., and Gruber, A.: ESA CCI Soil Moisture for improved Earth system
575 understanding: State-of-the art and future directions, Remote Sens. Environ., 203, 185-215,
576 2017.

Dudhia, J.: Numerical study of convection observed during the winter monsoon experiment using a mesoscale two-dimensional model, J. Atmos. Sci., 46, 3077-3107, 1989.

Ek, M., Mitchell, K., Lin, Y., Rogers, E., Grunmann, P., Koren, V., Gayno, G., and Tarpley, J.: Implementation of Noah land surface model advances in the National Centers for Environmental Prediction operational mesoscale Eta model, J. Geophys. Res. Atmos., 108, 2003.

Evans, J., Ward, H., Blake, J., Hewitt, E., Morrison, R., Fry, M., Ball, L., Doughty, L., Libre, J., and Hitt, O.: Soil water content in southern England derived from a cosmic-ray soil moisture observing system–COSMOS-UK, Hydrol. Processes, 30, 4987-4999, 2016.

Friesen, J., Rodgers, C., Oguntunde, P. G., Hendrickx, J. M., and van de Giesen, N.: Hydrotope-based protocol to determine average soil moisture over large areas for satellite calibration and validation with results from an observation campaign in the Volta Basin, West Africa, IEEE T Geosci Remote, 46, 1995-2004, 2008.

Fuamba, M., Branger, F., Braud, I., Batchabani, E., Sanzana, P., Sarrazin, B., and Jankowfsky, S.: Value of distributed water level and soil moisture data in the evaluation of a distributed hydrological model: Application to the PUMMA model in the Mercier catchment (6.6 km2) in France, J. Hydrol., 569, 753-770, 2019.

Gangopadhyay, S., Das Gupta, A., and Nachabe, M.: Evaluation of ground water monitoring network by principal component analysis, Groundwater, 39, 181-191, 2001.

Gilliland, E. K., and Rowe, C. M.: A comparison of cumulus parameterization schemes in the WRF model, Proceedings of the 87th AMS Annual Meeting & 21th Conference on Hydrology, 2007,

599 Hong, S.-Y., Noh, Y., and Dudhia, J.: A new vertical diffusion package with an explicit
600 treatment of entrainment processes, Mon. Weather Rev., 134, 2318-2341, 2006.

601 Jiménez, P. A., Dudhia, J., González-Rouco, J. F., Navarro, J., Montávez, J. P., and García-
602 Bustamante, E.: A revised scheme for the WRF surface layer formulation, Mon. Weather Rev.,
603 140, 898-918, 2012.

604 Kain, J. S.: The Kain-Fritsch convective parameterization: An update, J. Appl. Meteorol., 43,
605 http://dx.doi.org/10.1175/1520-0450(2004)043<0170:TKCPAU>2.0.CO;2, 2004.

606 Kodinariya, T. M., and Makwana, P. R.: Review on determining number of Cluster in K-Means
607 Clustering, Int. J. Adv. Res. Comput. Sci. Manag. Stud., 1, 90-95, 2013.

608 Laiolo, P., Gabellani, S., Campo, L., Silvestro, F., Delogu, F., Rudari, R., Pulvirenti, L., Boni,
609 G., Fascetti, F., and Pierdicca, N.: Impact of different satellite soil moisture products on the
610 predictions of a continuous distributed hydrological model, Int J Appl Earth Obs, 48, 131-145,
611 2016.

612 Lopez, P. L., Wanders, N., Schellekens, J., Renzullo, L. J., Sutanudjaja, E. H., and Bierkens,
613 M. F.: Improved large-scale hydrological modelling through the assimilation of streamflow
614 and downscaled satellite soil moisture observations, Hydrol. Earth Syst. Sci., 20, 3059-3076,
615 2016.

616 Mlawer, E. J., Taubman, S. J., Brown, P. D., Iacono, M. J., and Clough, S. A.: Radiative transfer
617 for inhomogeneous atmospheres: RRTM, a validated correlated-k model for the longwave,
618 Journal of Geophysical Research: Atmospheres, 102, 16663-16682, 1997.

619 Nash, J. E., and Sutcliffe, J. V.: River flow forecasting through conceptual models part I—A
620 discussion of principles, J. Hydrol., 10, 282-290, 1970.

621 Weather research and forcasting model, https://www.mmm.ucar.edu/weather-research-and-
622 forecasting-model, 2020.

623 Nistor, M. M.: Spatial distribution of climate indices in the Emilia-Romagna region, Meteorol.
624 Appl., 23, 304-313, 2016.

625 National Soil Moisture Network, http://nationalsoilmoisture.com/ , 2020.

626 Perry, M. A., and Niemann, J. D.: Analysis and estimation of soil moisture at the catchment
627 scale using EOFs, J. Hydrol, 334, 388-404, 2007.

628 Pistocchi, A., Calzolari, C., Malucelli, F., and Ungaro, F.: Soil sealing and flood risks in the
629 plains of Emilia-Romagna, Italy, J. Hydrol. Reg. Stud., 4, 398-409, 2015.

630 Rajib, M. A., Merwade, V., and Yu, Z.: Multi-objective calibration of a hydrologic model using
631 spatially distributed remotely sensed/in-situ soil moisture, J. Hydrol., 536, 192-207, 2016.

632 Robock, A., Vinnikov, K. Y., Srinivasan, G., Entin, J. K., Hollinger, S. E., Speranskaya, N. A.,
633 Liu, S., and Namkhai, A.: The global soil moisture data bank, Bull. Amer. Meteor. Soc., 81,
634 1281-1300, 2000.

635 Skamarock, W., Klemp, J., Dudhia, J., Gill, D., Barker, D., Duda, M., Huang, X., Wang, W.,
636 and Powers, J.: A description of the advanced research WRF Version 3, NCAR technical note,
637 Mesoscale and Microscale Meteorology Division, National Center for Atmospheric Research,
638 Boulder, Colorado, USA, 2008.

639 Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Barker, D. M., Wang, W., and Powers,
640 J. G.: A description of the advanced research WRF version 2, National Center for Atmospheric
641 Research Boulder, Colorado, USA, 2005.

642 Srivastava, P. K., Han, D., Rico-Ramirez, M. A., O'Neill, P., Islam, T., Gupta, M., and Dai, Q.:
643 Performance evaluation of WRF-Noah Land surface model estimated soil moisture for

644 hydrological application: Synergistic evaluation using SMOS retrieved soil moisture, J.
645 Hydrol., 529, 200-212, 2015.

646 Stéfanon, M., Drobinski, P., D'Andrea, F., Lebeaupin-Brossier, C., and Bastin, S.: Soil
647 moisture-temperature feedbacks at meso-scale during summer heat waves over Western
648 Europe, Clim. Dyn., 42, 1309-1324, 2014.

649 Thompson, G., Field, P. R., Rasmussen, R. M., and Hall, W. D.: Explicit forecasts of winter
650 precipitation using an improved bulk microphysics scheme. Part II: Implementation of a new
651 snow parameterization, Monthly Weather Review, 136, 5095-5115, 2008.

652 Uber, M., Vandervaere, J.-P., Zin, I., Braud, I., Heisterman, M., Legoût, C., Molinié, G., and
653 Nord, G.: How does initial soil moisture influence the hydrological response? A case study
654 from southern France, Hydrol. Earth Syst. Sci., 22, 6127-6146, 2018.

655 Vachaud, G., Passerat de Silans, A., Balabanis, P., and Vauclin, M.: Temporal stability of
656 spatially measured soil water probability density function 1, Soil Sci. Soc. Am. J., 49, 822-828,
657 1985.

658 Vereecken, H., Huisman, J., Bogena, H., Vanderborght, J., Vrugt, J., and Hopmans, J. W.: On
659 the value of soil moisture measurements in vadose zone hydrology: A review, Water Resour.
660 Res., 44, 2008.

661 Western, A. W., Zhou, S.-L., Grayson, R. B., McMahon, T. A., Blöschl, G., and Wilson, D. J.:
662 Spatial correlation of soil moisture in small catchments and its relationship to dominant spatial
663 hydrological processes, J. Hydrol., 286, 113-134, 2004.

664 Wold, S., Esbensen, K., and Geladi, P.: Principal component analysis, Chemometr Intell Lab
665 Syst 2, 37-52, 1987.

Zaidi, S. M., and Gisen, J. I. A.: Evaluation of Weather Research and Forecasting (WRF) Microphysics single moment class-3 and class-6 in Precipitation Forecast, MATEC Web of Conferences, 2018, 03007,

Zaitchik, B. F., Santanello, J. A., Kumar, S. V., and Peters-Lidard, C. D.: Representation of soil moisture feedbacks during drought in NASA unified WRF (NU-WRF), J. Hydrometeorol., 14, 360-367, 2013.

Zhuo, L., Dai, Q., Islam, T., and Han, D.: Error distribution modelling of satellite soil moisture measurements for hydrological applications, Hydrol. Process., 30, 2223-2236, 2016.

Zhuo, L., and Han, D.: Multi-source hydrological soil moisture state estimation using data fusion optimisation, Hydrol. Earth Syst. Sci., 21, 3267-3285, 2017.

Zhuo, L., Dai, Q., Han, D., Chen, N., and Zhao, B.: Assessment of simulated soil moisture from WRF Noah, Noah-MP, and CLM land surface schemes for landslide hazard application, Hydrol. Earth Syst. Sci., 23, 4199-4218, 2019a.

Zhuo, L., Dai, Q., Han, D., Chen, N., Zhao, B., and Berti, M.: Evaluation of Remotely Sensed Soil Moisture for Landslide Hazard Assessment, IEEE J-STARS, 12, 162-173, 2019b.

Zreda, M., Shuttleworth, W. J., Zeng, X., Zweck, C., Desilets, D., Franz, T., and Rosolem, R.: COSMOS: the cosmic-ray soil moisture observing system, Hydrol. Earth Syst. Sci., 16, 2012.

Zwieback, S., Westermann, S., Langer, M., Boike, J., Marsh, P., and Berg, A.: Improving permafrost modeling by assimilating remotely sensed soil moisture, Water Resour. Res., 55, 1814-1832, 2019.

687

688 **Table 1.** WRF parameterizations used in this study.

	Settings/ Parameterizations	References
Map projection	Lambert	
Central point of domain	Latitude: 44.54; Longitude: 11.02	
Latitudinal grid length	5 km	
Longitudinal grid length	5 km	
Model output time step	Daily	
Nesting	Two-way	
Land surface model	Noah-MP	
Simulation period	1/1/2006 – 31/12/2015	
Spin-up period	1/1/2005 – 31/12/2005	
Microphysics	New Thompson	(Thompson et al., 2008)
Shortwave radiation	Dudhia scheme	(Dudhia, 1989)
Longwave radiation	Rapid Radiative Transfer Model	(Mlawer et al., 1997)
Surface layer	Revised MM5	(Jiménez et al., 2012;Chen and Dudhia, 2001)
Planetary boundary layer	Yonsei University method	(Hong et al., 2006)
Cumulus Parameterization	Kain-Fritsch (new Eta) scheme	(Kain, 2004)

689

690

691 **Table 2.** The relationship between the percentage of grids, and the cross-correlation.

Cross-correlation (*r*)	Percentage of grids (%)
0.5	85
0.6	78
0.7	70
0.8	52
0.9	15
0.95	3

692

693 **Table 3.** The number of components and grids to reach % variance threshold (based on the
694 PCA method and the elbow curve method).

Variance (%)	Components	Number of grids
70.0	1	6
75.0	1	7
80.0	2	11
85.0	2	21
90.0	3	50
92.5	3	94
95.0	3	184
97.5	3	367

695

696

697

698

Table 4. *NSE* and correlation *r* performance of CA_Med and CA_Max.

Variance	No. of grids	CA_Max_Mean		CA_Med_Mean		CA_Max_STD		CA_Med_STD	
		NSE	r	NSE	r	NSE	r	NSE	r
70.0	6	0.831	0.978	0.949	0.985	0.601	0.834	0.716	0.876
75.0	7	0.851	0.984	0.978	0.993	0.778	0.887	0.746	0.870
80.0	11	0.894	0.990	0.991	0.996	0.867	0.945	0.901	0.951
85.0	21	0.976	0.997	0.991	0.998	0.926	0.967	0.930	0.976
90.0	50	0.988	0.998	**0.995**	**0.999**	0.963	0.986	**0.973**	**0.990**
92.5	94	0.997	0.998	0.990	0.999	0.969	0.989	0.960	0.992
95.0	184	0.994	0.999	0.985	0.999	0.932	0.990	0.914	0.986
97.5	367	0.988	1.000	0.983	1.000	0.910	0.986	0.895	0.982

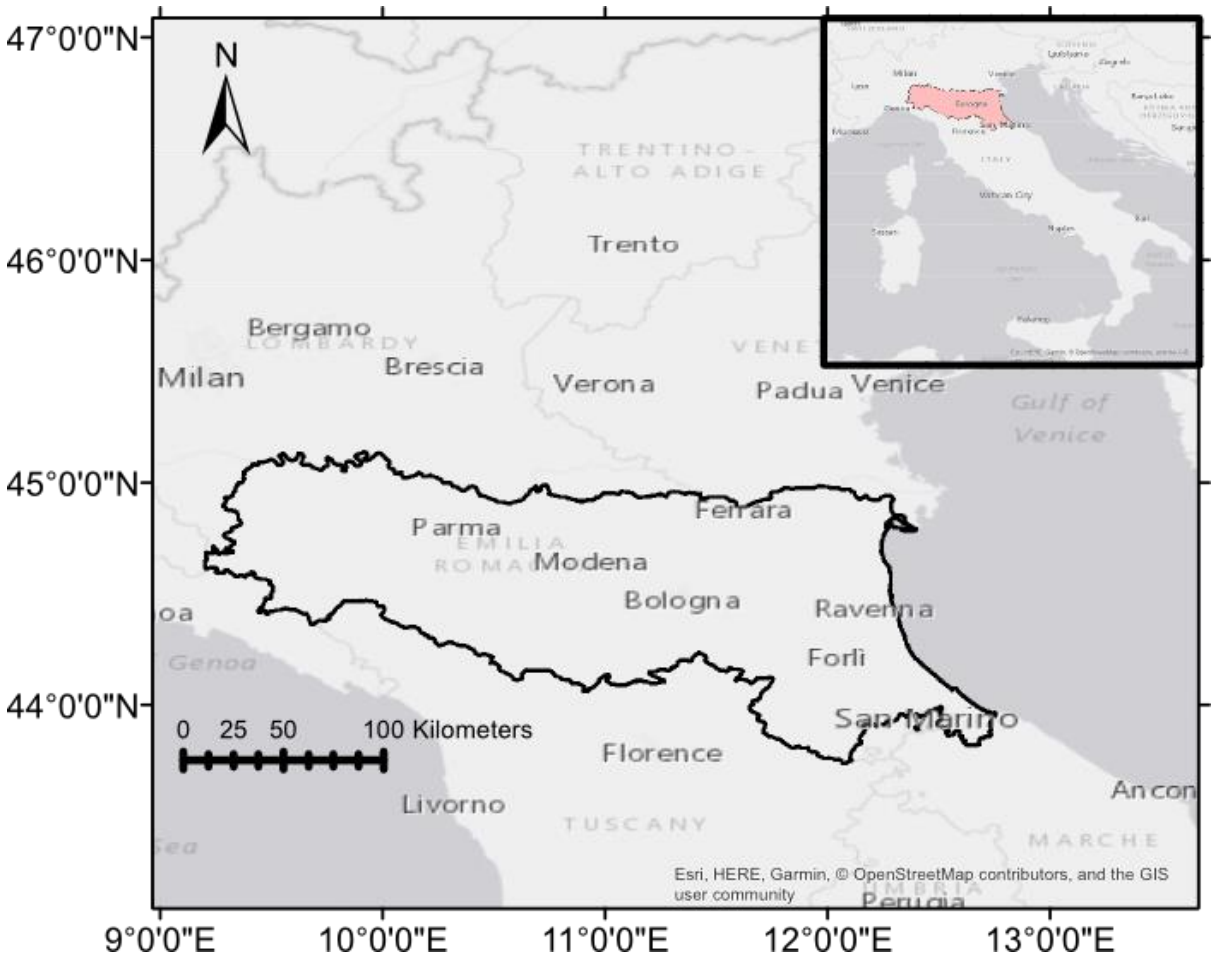


Figure 1. The geographical map of the Emilia Romagna region. The copyright of the background map belongs to Esri (Light Gray Canvas Basemap).

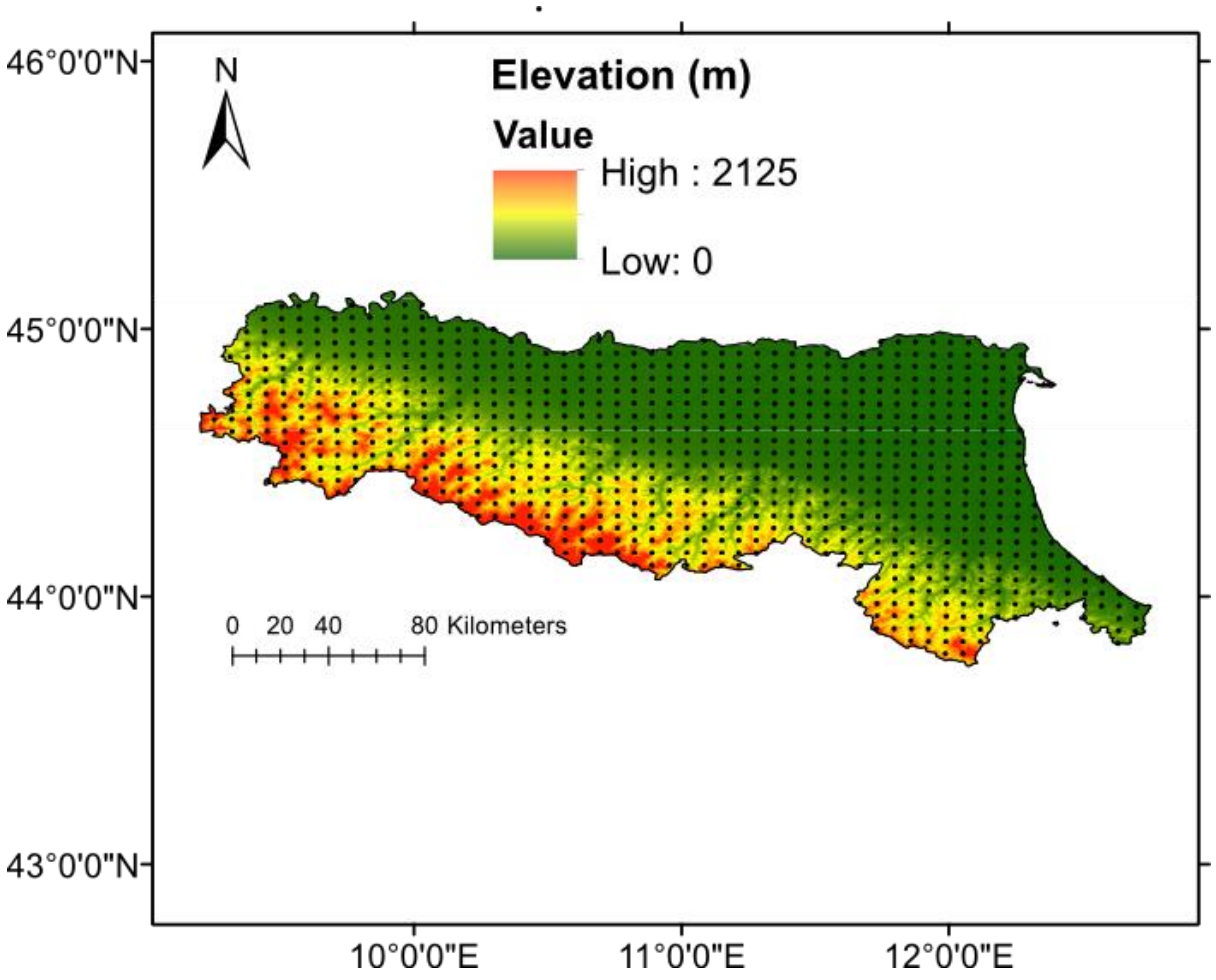


Figure 2. WRF grids used in the analysis, with DEM map in the background.

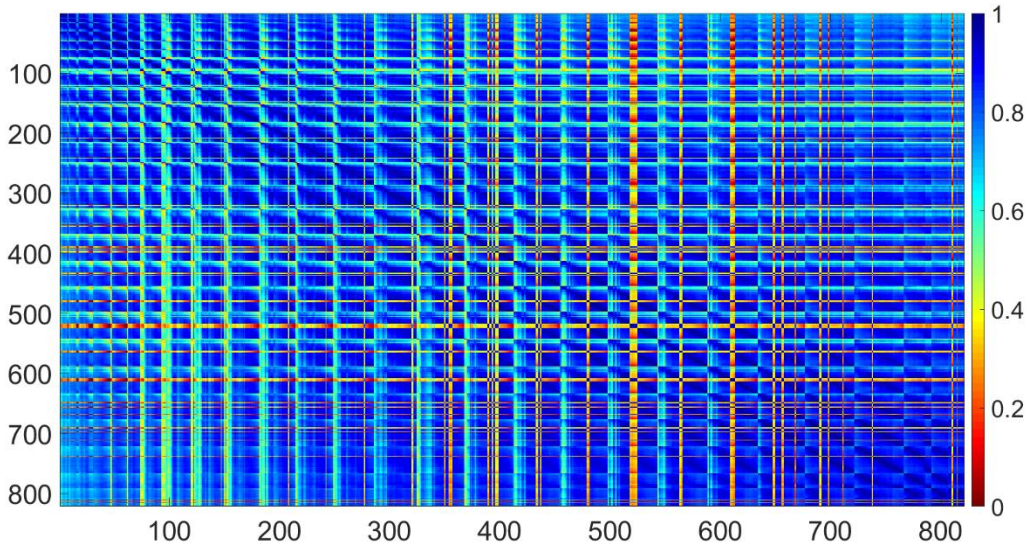


Figure 3. Cross correlation matrix for the whole catchment.

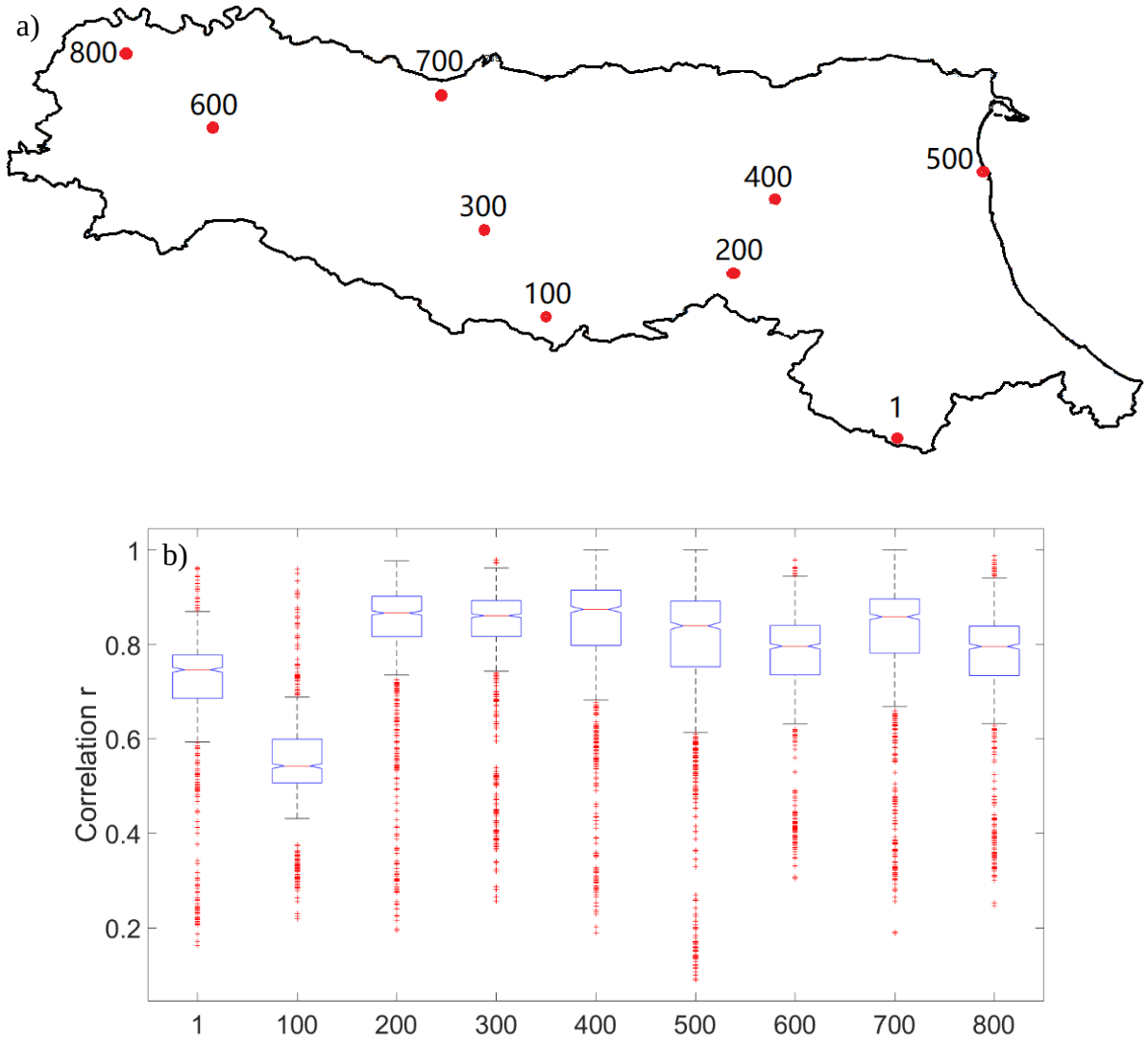
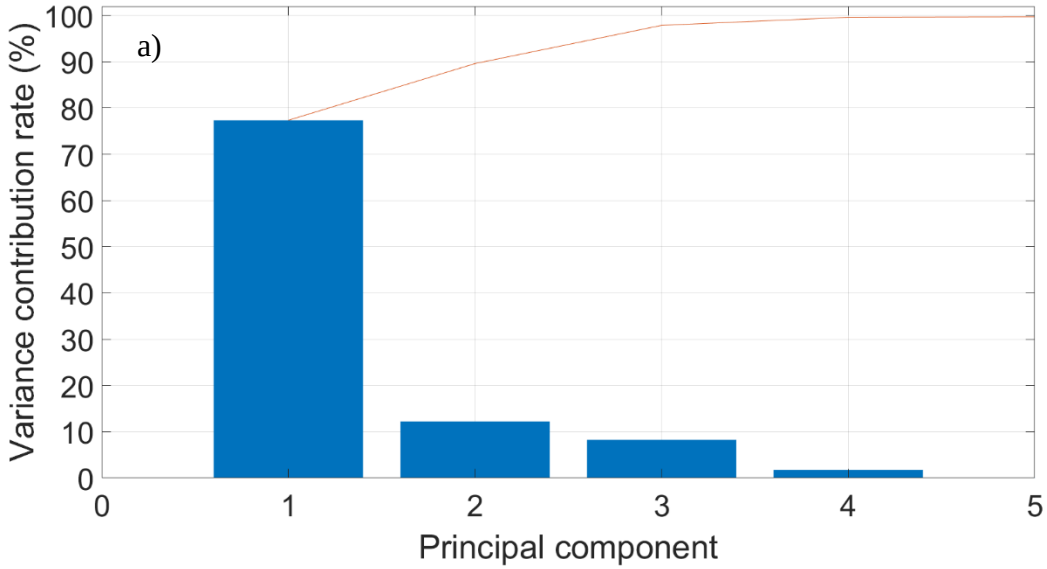
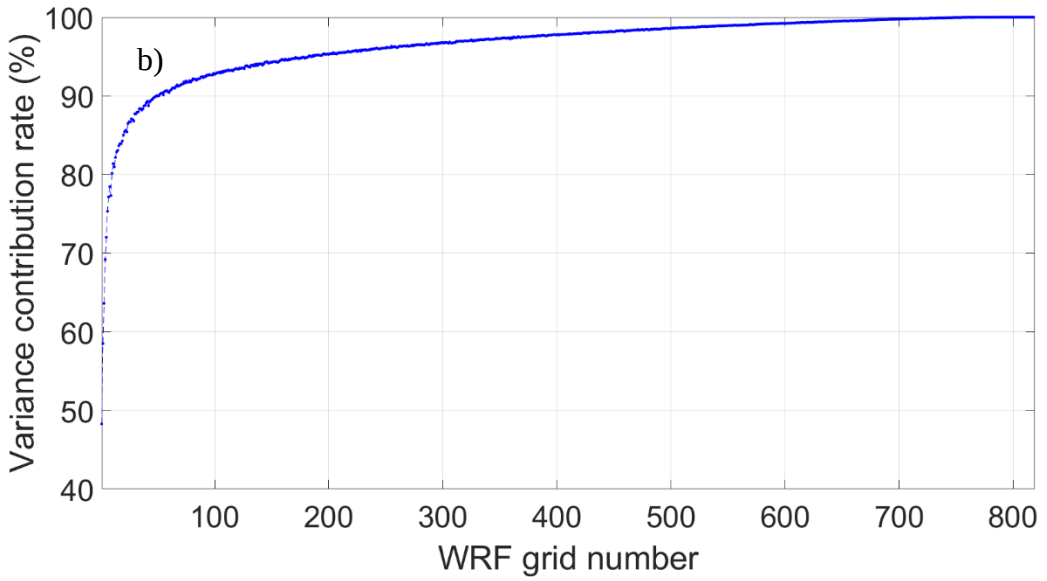


Figure 4. a) WRF selected grid number; b) correlation boxplot for the selected grids as highlighted in red in a). For the boxplot, it shows the minimum, maximum, 0.25, 0.50, and 0.75 percentiles and outliers (red cross).



732



733

734 **Figure 5.** a) PCA analysis; b) elbow curve.

735

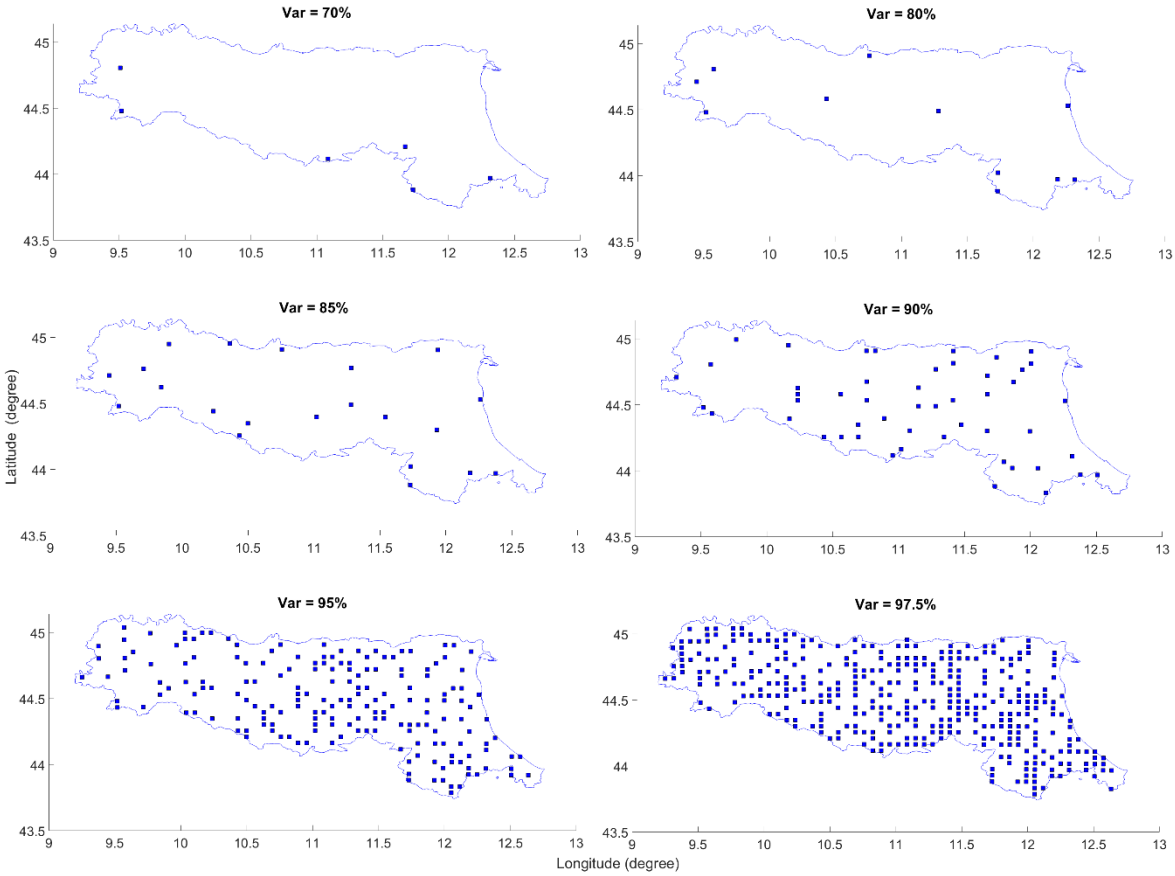
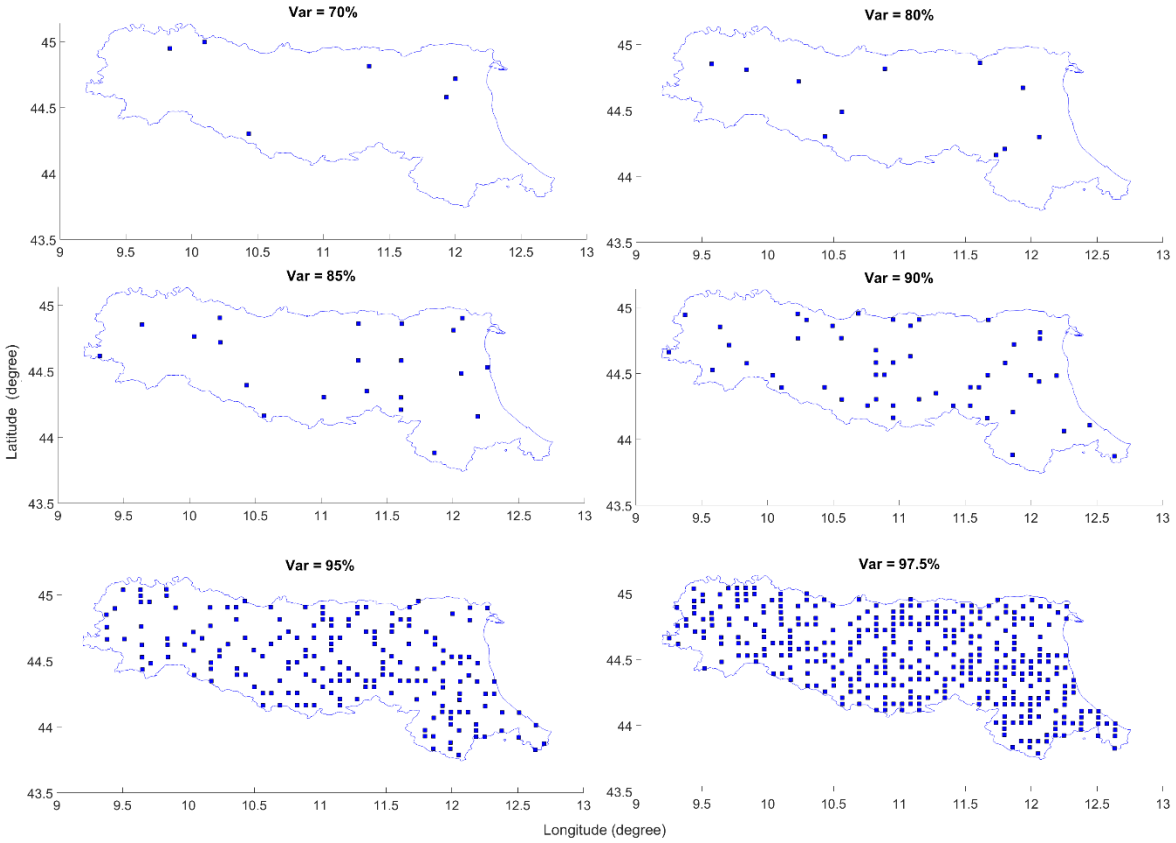


Figure 6. Designed soil moisture sensor locations, based on CA-Max. The total number of grids used for the design is 6, 11, 21, 50, 184, 367 for 70%, 80%, 85%, 90%, 95%, 97.5% variance, respectively.



741

742 **Figure 7.** Designed soil moisture sensor locations, based on CA-Med. The number of grids
743 used for the design is 6, 11, 21, 50, 184, 367 for 70%, 80%, 85%, 90%, 95%, 97.5% variance,
744 respectively.

745

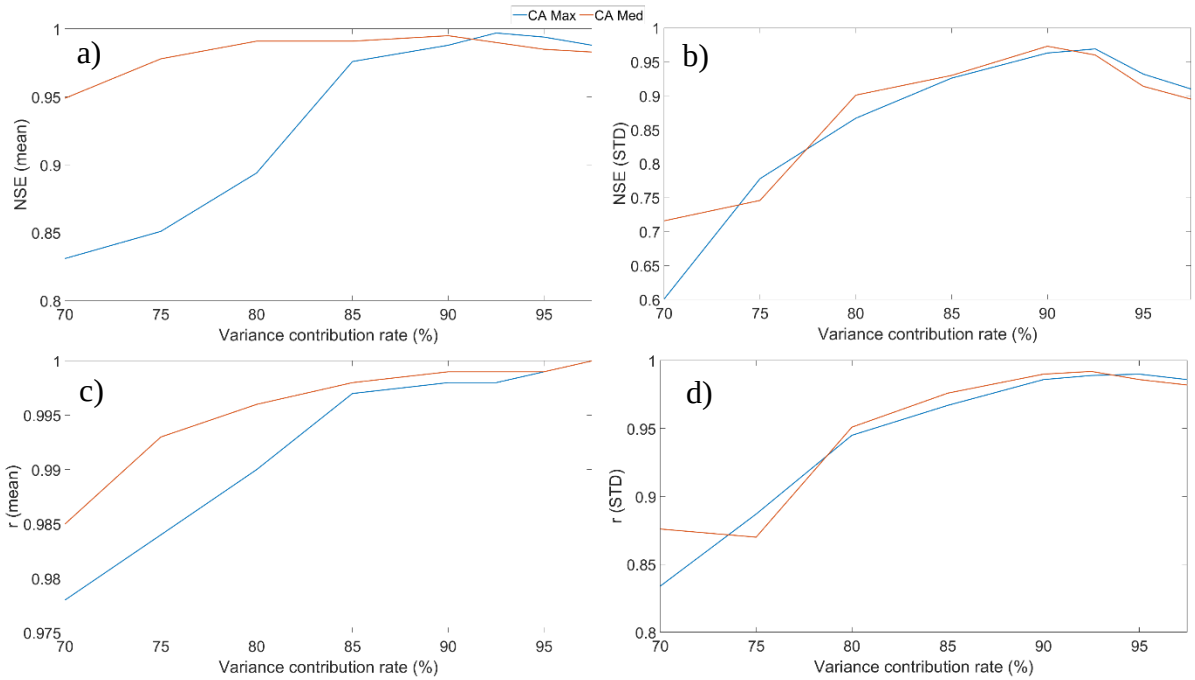
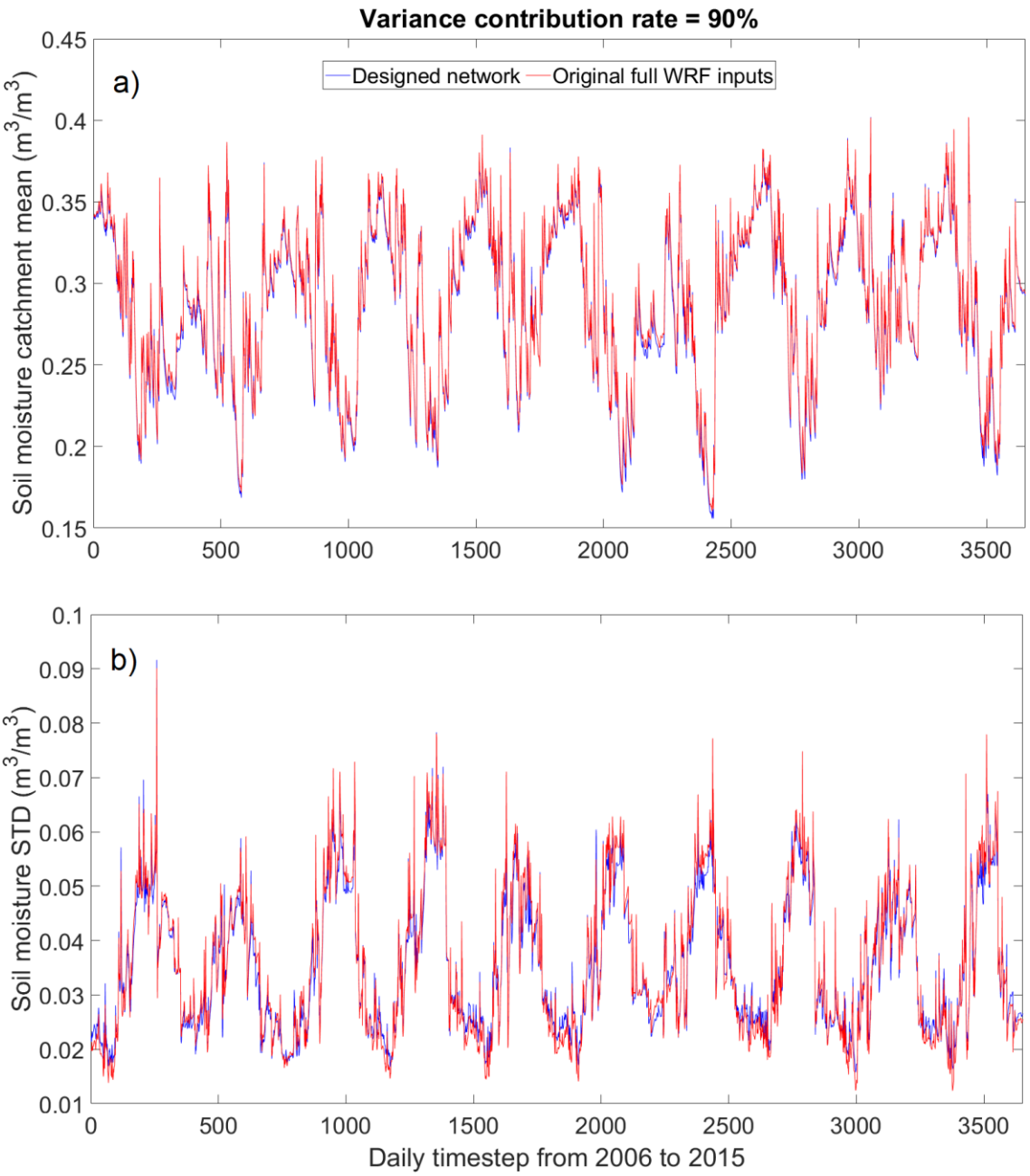


Figure 8. *NSE* and *r* plots: a) *NSE* performance based on the areal mean soil moisture, b) *NSE* performance based on the areal standard deviation soil moisture (STD), c) *r* performance based on the areal mean soil moisture, d) *r* performance based on the areal standard deviation soil moisture.



752

753 **Figure 9.** a) The areal mean soil moisture of the designed and the WRF's full-input networks,
754 b) the areal soil moisture standard deviation of the designed and the WRF's full-input networks.
755 The designed network is based on CA-Med, 90% variance contribution rate, and 50 sensors.

756

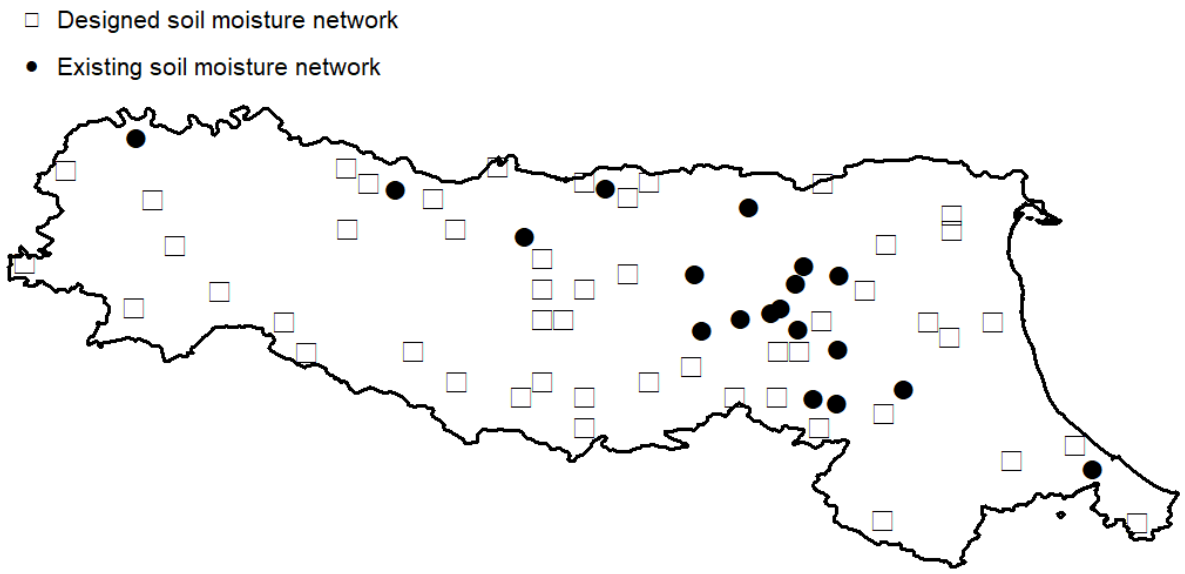
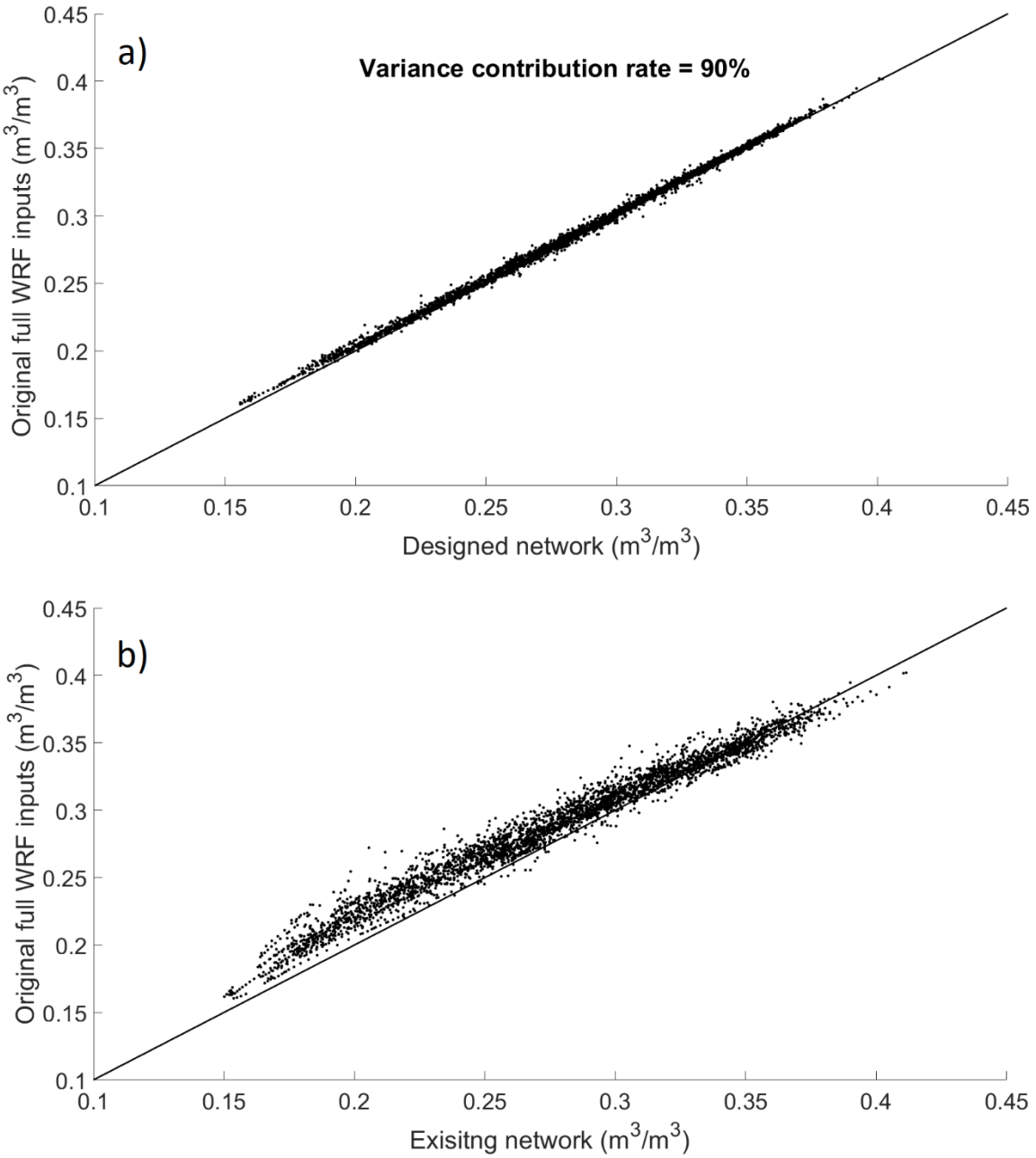


Figure 10. Comparison between the existing and the designed soil moisture networks.



760

761 **Figure 11.** Scatterplots for areal mean soil moisture: a) WRF full grid inputs against the
762 proposed network (*NSE* = 0.995, *r* = 0.998); b) WRF full grid inputs against the existing in-
763 situ network (*NSE* = 0.889, *r* = 0.987).

764

765

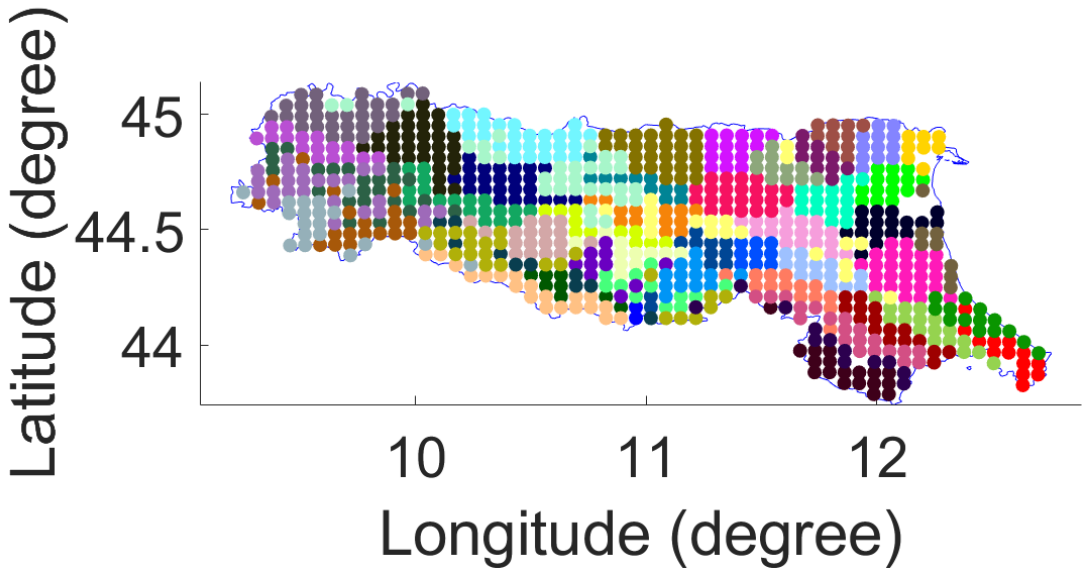


Figure 12. Cluster map.