Remote Sensing of Environment 258 (2021) 112377

Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

Comparison of gap-filling techniques applied to the CCI soil moisture database in Southern Europe

Laura Almendra-Martín [a,*], José Martínez-Fernández [a], María Piles [b], Ángel González-Zamora [a]

[a] Instituto Hispano Luso de Investigaciones Agrarias, CIALE, Universidad de Salamanca, Villamayor, Spain
[b] Image Processing Laboratory, Universitat de València, 46980, Valencia, Spain

ARTICLE INFO

Keywords:
Gap-filling
Soil moisture
CCI
Support vector machines

ABSTRACT

Soil moisture (SM) is a key variable that plays an important role in land-atmosphere interactions. Monitoring SM is crucial for many applications and can help to determine the impact of climate change. Therefore, it is essential to have continuous and long-term databases for this variable. Satellite missions have contributed to this; however, the continuity of the series is compromised due to the data gaps derived by different factors, including revisit time, presence of seasonal ice or Radio Frequency Interference (RFI) contamination. In this work, the applicability of different gap-filling techniques is evaluated on the ESA Climate Change Initiative (CCI) SM combined product, which is the longest available satellite-based SM data record. The methods used were linear, cubic and autoregressive interpolation and support vector machines (SVMs). This study focused on Southern Europe and spanned the years 2003–2015. The different methods were applied in the temporal and spatial domains and evaluated using the holdout cross-validation technique. A set of variables was introduced in the SVM model to estimate SM, namely, land surface temperature, precipitation, normalized difference vegetation index (NDVI), potential evaporation, soil texture and geographical coordinates. For the SVMs, several combinations of these variables were considered, including a principal component analysis (PCA) containing all of them. Although the different methods show a generally good performance, the SVM method outperforms the rest. Using the SM of the precedent day (SM_{t-1}) is key to obtain good estimates. The median value of the correlation coefficient (R) obtained with the SVM and the SM_{t-1} series in the temporal analysis was 0.83, and the RMSE was 0.025 m^3m^{-3}. Similar results were obtained in the spatial analysis, with the best performance ($R = 0.88$; RMSE $= 0.024$ m^3m^{-3}) obtained by the SVM using the SM_{t-1} series and the static variables. The application of PCA to input variables was not beneficial, and the interpolation methods failed when dealing with large spatial or temporal gaps. A validation of the CCI SM series with *in situ* SM data from four networks located in Spain, France, Germany and Italy was also performed and no substantial differences were observed between results obtained with the original and with the reconstructed series. In addition, best inputs obtained with SVM were used to evaluate the random forest (RF) method in the temporal and spatial domain. This method showed a good ability to estimate soil moisture values in the temporal domain but to a lesser extent than SVM while for the spatial domain it did not seem to be as accurate. Our results confirm that we can efficiently deal with spatio-temporal gaps on observational SM databases using the SVM method and the past time series and soil texture as supporting information.

1. Introduction

Soil moisture (SM) is a relevant variable in land-atmosphere interactions as it controls the water, energy and carbon cycles, behaves as storage for precipitation, governs runoff and limits plant transpiration (Seneviratne et al., 2010). It is also a crucial variable in agricultural applications (Champagne et al., 2019) and many environmental studies, such as flood forecasting (Brocca et al., 2011), drought monitoring (Liu et al., 2019; Martínez-Fernández et al., 2016) and evaporation modelling (Miralles et al., 2011). Given its importance within the Earth system, SM was listed as one of the 50 essential climate variables (ECVs) by the Global Climate Observing System (GCOS, 2010) in 2010, and many efforts have been dedicated to the global mapping of SM in recent decades. Several SM products have been developed and validated with

* Corresponding author at: Instituto Hispano Luso de Investigaciones Agrarias, CIALE, Universidad de Salamanca, Duero, 12, 37185 Villamayor, Spain.
E-mail address: lauraalmendra@usal.es (L. Almendra-Martín).

https://doi.org/10.1016/j.rse.2021.112377
Received 10 July 2020; Received in revised form 20 January 2021; Accepted 23 February 2021
Available online 3 March 2021

different objectives, characteristics and data sources (Beck et al., 2020). Examples include the International Soil Moisture Network (ISMN) (Dorigo et al., 2011), which consists of soil moisture series from *in situ* networks from all over the world. This soil moisture database is extremely useful, but has obvious limitations in terms of spatial coverage. Other approaches are the microwave active or passive satellite missions (*e.g.*, Soil Moisture Ocean Salinity, SMOS, and Soil Moisture Active and Passive, SMAP) that provide a continuous spatio-temporal monitoring of this variable. The estimations are based on the high sensitivity of the brightness temperature measured by passive sensors, or the radar backscattering coefficient measured by active sensors, to the dielectric constant of soil, which is directly related to the soil moisture content. However, the retrieval of the variable can be complex due to various factors such as dense vegetation (Jackson, 1993). Furthermore, by merging different soil moisture estimations from different remote sensing missions, multi-satellite databases can be obtained (*e.g.*, Soil Moisture Operational Products System, SMOPS, and Climate Change Initiative, CCI). These kind of products present the advantage of having a better spatio-temporal coverage than those obtained with just one sensor. Lastly, land surface models and reanalysis products (*e.g.*, ERA5-Land and Global Land Data Assimilation System, GLDAS) provide a complete spatio-temporal coverage, since soil moisture is estimated by mathematical models incorporating also other variables.

The CCI programme from the European Space Agency (ESA) aims to provide long-term observational datasets of biogeophysical variables by taking advantage of the satellite measurements acquired during the observational era, *i.e.*, from 1970's to the present (https://www.esa-soilmoisture-cci.org, Soil Moisture Climate Change Initiative (CCI), 2020). These variables are all integrated in the Copernicus Climate Service (C3S), a Copernicus Earth Observation Programme service focused on climate research that provides climate information and data (https://climate.copernicus.eu, Copernicus Climate Service (C3S), 2020). SM is one of these variables (hereafter CCI SM). The CCI SM product integrates SM products from four active and seven passive microwave satellite sensors, making the largest available existing observational SM data record (Dorigo et al., 2017; Gruber et al., 2019; Gruber et al., 2017). This database has been extensively validated in different regions of the world including Southern Europe (Al-Yaari et al., 2019; González-Zamora et al., 2019; An et al., 2016; Dorigo et al., 2015; Ikonen et al., 2018; McNallya et al., 2016). All these studies found great accordance between the CCI product and the different *in situ* or reanalysis soil moisture series, although with some uncertainties, especially for the first years of the period when the data series have more gaps. Furthermore, CCI is extensively used for different applications, such as the study of global trends in SM (Feng, 2016; Qiu et al., 2016; Wang et al., 2016), precipitation estimations (Ciabatta et al., 2018), crop models (Sakai et al., 2016), the relationship between drought and climatic variables (Nicolai-Shaw et al., 2017), the assessment of the impact of El Niño drought conditions (Dorigo et al., 2016) or even tree growth tracking (Martínez-Fernández et al., 2019).

The ESA CCI SM product quality has steadily increased with each successive release, and the merged products generally outperform the single-sensor input products (Dorigo et al., 2017). Nevertheless, this database poses some limitations for several applications. For example, higher spatial resolutions are required to serve regional applications. This problem has been addressed in some studies that applied downscaling techniques to microwave-based SM products (Mascaro et al., 2011; Peng et al., 2017; Piles et al., 2016; Piles et al., 2014; Srivastava et al., 2013). Additionally, many studies have problems dealing with the spatio-temporal gaps in the data, which are caused by a variety of factors (*e.g.*, RFI contamination, different satellite revisit times, presence of ice or snow, high uncertainty of retrievals in coastal and mountain areas). Several methods have been proposed in the literature to overcome this non-uniform effective sampling of SM observational data at different spatial and temporal scales. For *in situ* SM databases, some studies have compared a suite of different gap-filling methods (Dumedah et al., 2014; Dumedah and Coulibaly, 2011; Ford and Quiring, 2014; Kornelsen and Coulibaly, 2014). Regarding satellite images, Zhang and Chen (2016) proposed the satellite and *in situ* sensor collaborated reconstruction (SICR) method for filling Gaofeng-1 SM gaps. They classified the missing pixels based on their characteristics and their similarity or proximity to the *in situ* data and established four rules for the reconstruction based on linear regression or ordinary kriging. Xing et al. (2017) improved the SICR method by applying machine-learning techniques. While the two methodologies offer good results, they can be applied only to regions where *in situ* SM data exist. Xiao et al. (2016) proposed a way to efficiently reconstruct the satellite series by using the GLDAS Noah model but only in one-year period. In that study, satellite data were used to estimate the model control variables, and meteorological data were incorporated to force the model to simulate the temporal dynamics. Wang et al. (2012) applied a penalized least square method based on three-dimensional discrete cosine transformation to the Advanced Microwave Scanning Radiometer - Earth Observing System (AMSR-E) SM product, which was originally proposed by García (2010) for smoothing multidimensional data with missing values. This methodology enables the simultaneous consideration of spatial and temporal database information but leads to poor predictions when the spatial differences are large. Furthermore, Cui et al. (2016) proposed a reconstruction of the Fengyun-3B Microwave Radiation Imager (FY-3B/MWRI) SM product using Moderate Resolution Imaging Spectroradiometer (MODIS) products by applying artificial neural networks (ANNs). This method was able to capture the SM dynamics, but they found uncertainties regarding the algorithm, which should be improved in the freezing-thaw period.

Despite the previous research aimed at completing SM databases, only a few have been applied to the challenging case of the long-term multi-satellite CCI SM product. In Llamas et al. (2020), three spatial methodologies were evaluated over a region in the U.S. Midwest: ordinary kriging, regression kriging and general linear models (GLMs). These methods are based on the spatial distribution of SM or its relationship with other variables (temperature and precipitation) and show good performance, especially the two kriging approaches. Cui et al. (2019) applied a modified algorithm from Cui et al. (2016) to the CCI SM product in the Tibetan Plateau. This method was based on a general regression neural network (GRNN) and used the SM, land surface temperature (LST), normalized difference vegetation index (NDVI), albedo and digital elevation model (DEM) as inputs. Liu et al. (2020) proposed the use of SMAP data to complete the CCI gaps. While the resampled SMAP SM series proved to be able to reasonably fill CCI SM gaps, the complete series reconstruction depends upon the availability of SMAP data.

Studies comparing the performance of different gap-filling methodologies to SM databases are limited, and they only partially address the specific case of long-term multi-satellite observational SM series. In this study we aim to bridge this gap by analysing the performance of gap-filling methodologies with different levels of complexity to the ESA CCI SM product: from simple ones, such as linear interpolation, to more sophisticated ones, such as those based on machine learning (ML). We selected ML techniques since they can integrate multivariate information and have been shown to excel in a variety of SM applications, from nonparametric and nonlinear classification to regression techniques (Lary et al., 2016). Specifically, the support vector machines (SVMs) for regression problems allow good generalization even with small training datasets (Ali et al., 2015; Mountrakis et al., 2011) and, in some cases, work better than ANNs (Ahmad et al., 2010). The effectiveness of SVMs to fill temporal gaps in ground-based observation databases has been proven elsewhere (Gill et al., 2006), but they have not been previously used to complete satellite databases.

Knowledge of the SM spatial distribution and its dynamics, anomalies and trends across time is fundamental to assess and quantify the impact of climate change on the water cycle. This highlights the need to fill the data gaps and improve the temporal sampling and observation density of the current observational soil moisture databases spanning

the last 40 years. This work evaluates a suite of gap-filling methods of varying complexities for the case of long-term satellite-based SM databases (the CCI SM). This research is focused on the Southern part of Europe and covers spatial and temporal domains. Among the wide range of possible features that could explain the SM variability, the LST, NDVI, precipitation, potential evaporation and soil texture were chosen and used as inputs for the SVM models. The remainder of this article is organized as follows. Section 2.1 provides a description of the databases used in this study; Section 2.2 introduces the methods, the developed models and the overall validation strategy. Section 3 shows the results obtained by the different methods in the temporal and spatial domains and discusses the best results obtained in each analysis and their applicability. Conclusions and perspectives from this article are provided in Section 4.

2. Material and methods

2.1. Datasets

The CCI SM product is based on a merging algorithm that harmonizes the SM retrievals from available active and passive microwave sensors. The single-sensor SM products are merged into three different products depending on the type of sensors used: the active, the passive and the combination of the two. All products are provided in global daily maps in a regular grid of 0.25° (Dorigo et al., 2017). This merging algorithm has been updated and improved in the different versions of the product, which incorporate an increasing number of microwave sensors (Gruber et al., 2019; Gruber et al., 2017). In this study, the combined product of the latest version (v4.5) was used, and only the data with best quality were considered. This was achieved by screening out data with reported inconsistencies by using the quality flag variable. The CCI SM product also provides SM uncertainty values that are used to interpret the results. The combined product covers a 40-year period (November 1978 to December 2018) and includes eleven microwave sensors, seven of which are passive and four of which are active (Fig. 1), with different technical characteristics and coverage periods (Gruber et al., 2019). The availability of a number of sensors for specific time periods leads to important differences in spatiotemporal coverage, resulting in periods when a single operating sensor was used to retrieve global SM but also in periods when up to five simultaneous estimations were merged. Due to this fact, the spatiotemporal distribution of data gaps is highly heterogeneous.

For the study area, the south of Europe, the percentage of data availability did not exceeded 20% in the first 20 years of the series (Fig. 1). The amount of available data significantly increased notably to 40–50% in 2002–2003, when AMSR-E was added. Therefore, two periods can be clearly distinguished in the series in terms of data availability, with the tipping point being 2003. For this reason and due to the limited availability of complementary databases in the first period, this study focuses on the second period, *i.e.*, 2003–2015. In this period, there was an increasing trend in data coverage over time until 2012, when AMSR-E operations ceased. The percentage of available data increased again in 2013 with AMSR-2 data, reaching a plateau of approximately 80% lasting until the end of the series. The mean percentage of SM data over Southern Europe for the study period is 62%, but its spatial distribution is non-homogeneous; some coastal pixels and mountain regions are the ones with the lowest percentage of data (see Fig. 2a). But the majority of the area of study has between 70 and 80% of available data (Fig. 2b).The temporal distribution and the gap length are also irregular. While the most common gaps last one or two days, there are also gaps of more than one year (Fig. 2c).

Several atmospheric, geophysical and hydrological variables are related to SM and can help capture its variability across space and time (Korres et al., 2013; Sandholt et al., 2002; Wang et al., 2007, 2017). In this study, the LST, NDVI, precipitation (P), potential evaporation (E_p) and soil texture were used to estimate the missing values of the CCI SM database (Table 1). The databases of these variables were chosen with the criteria of providing the longest temporal coverage and the fewest data gaps. Furthermore, remote sensing estimations rather than models were chosen when possible to avoid including additional uncertainties in the SM estimations. In addition, values were filtered using the provided quality flags to keep only the pixels with the highest quality. Due to the spatial resolution differences among the products used, all databases were projected to the WGS84 coordinate system and resampled into a common 0.25° grid by averaging the pixel values. This is a

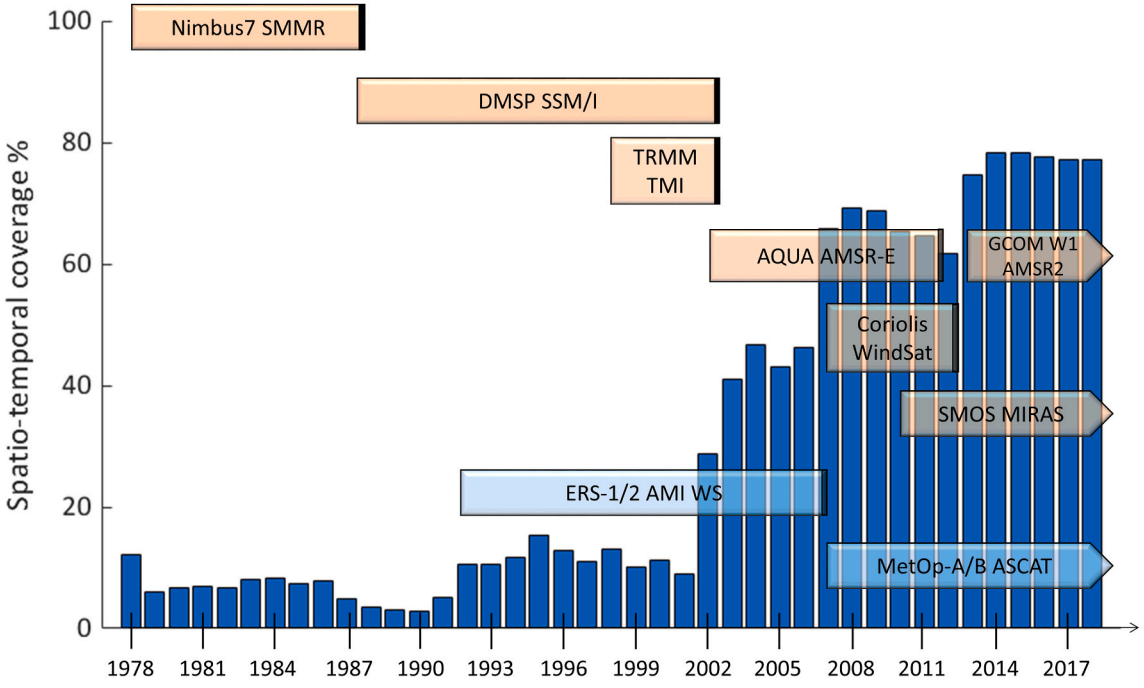

Fig. 1. Timeline of the passive (orange) and active (blue) microwave sensors that generate the CCI SM product for version v4.5 and its annual percentage of available data in Southern Europe. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

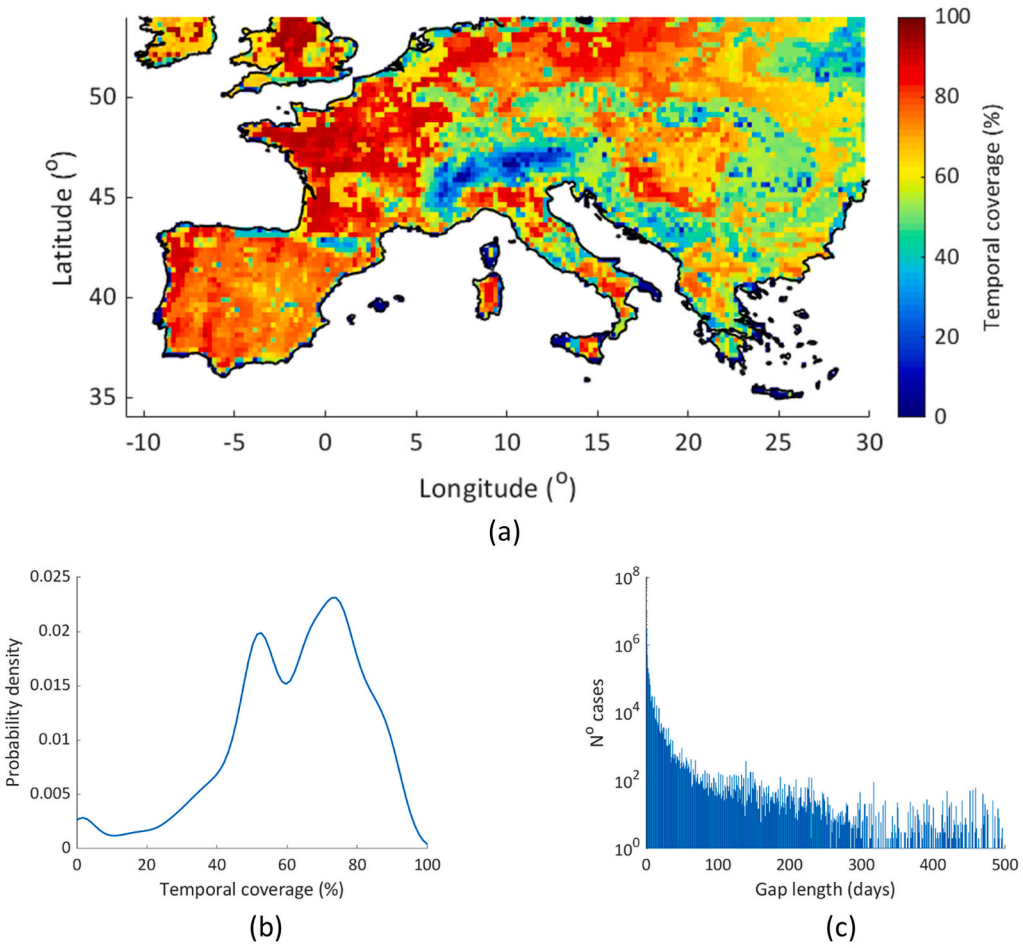

Fig. 2. Temporal coverage of data available for the study period (2003–2015) (a). Its probability density function (b). Histogram of the data gap lengths for the period (2003–2015) in Southern Europe (c).

Table 1
Characteristics of the datasets used in the study.

Variable	Units	Product	Spatial resolution	Temporal coverage	Temporal resolution	Reference
SM	m^3m^{-3}	CCI SM combined v4.5	0.25°	1978/11/01–2018/12/31	1 d	(Dorigo et al., 2017; Gruber et al., 2017, 2019)
LST	K	LSA-001 CM SAF	0.05°	1991/01/01–2015/12/31	1 h	(Duguay-Tetzlaff et al., 2017)
P	mm	GPM IMERG Final Precipitation L3 v06	0.1°	2000/06/01 - Present	1 d	(Huffman et al., 2019)
NDVI	–	MOD13A2 v6	1 km	2000/02/18 - Present	16 d	(Didan, 2015)
E_p	mm/d	GLEAM ET v3.3 b	0.25°	2003/01/01–2018/09/30	1 d	(Martens et al., 2017; Miralles et al., 2011)
Soil texture	%	European Soil Data Centre (ESDAC) LUCAS topsoil	500 m	–	–	(Ballabio et al., 2016)

common and well-accepted practice when dealing with coarse resolution remote sensing data (Liu et al., 2020; Sandholt et al., 2002; Qu et al., 2019). In addition, the 16-day NDVI product was interpolated and the hourly LST product was averaged to obtain daily series. *In situ* SM data were also used to validate the CCI SM reconstructed series following the methodology of González-Zamora et al. (2019). For this, four SM networks of the ISMN over the south of Europe were used: REMEDHUS network (González-Zamora et al., 2016) located in Spain, the TERENO network (Zacharias et al., 2011) in Germany, the UMBRIA network (Brocca et al., 2008) in Italy, and the ORACLE network (Tallec et al., 2015) in France.

2.2. Gap-filling techniques

The gap-filling techniques chosen for this study were (i) linear interpolation, (ii) cubic interpolation, (iii) SVMs and (iv) SVMs combined with PCA. Furthermore, in the SVMs, an analysis of different combinations of input variables was carried out. The combination of inputs with the best accuracy was used to evaluate the random forest (RF) method. With the aim of evaluating the performance of the different methods, a holdout cross-validation was performed with nine replicates (Browne, 2000; Pérez-Planells et al., 2015). SM values were separated into the training and the test subsets (70% and 30% of the existing values, respectively). The partition of the datasets randomly reproduced the spatiotemporal distribution of the CCI SM gaps of the

study area and is the same for all the methods to ensure robustness and consistency in the inter-comparisons. Once the reconstructed SM series were obtained from the training set, the new values were validated with the test set. The statistics calculated for the validation assessment with the test set were the Pearson correlation coefficient (R), the average error bias, the root mean square error (RMSE) and the centered RMSE (cRMSE). The last three are expressed in volumetric units (m^3m^{-3}). These metrics are commonly used in satellite SM validation exercises (Entekhabi et al., 2010).

2.2.1. Interpolation methods

For the spatial domain, the objective was to complete each daily map of the CCI SM series. Delaunay triangulation-based 2-D linear interpolation (LI) and cubic interpolation (CI) were carried out to achieve this goal. Additionally, the missing values in the temporal domain (*i.e.*, from each pixel time series) were estimated by using an LI and a CI by splines. This last algorithm preserves the monotony of the data in an interval by using a cubic function (Fritsch and Carlson, 1980).

For the temporal domain, an autoregressive (AR) model was used. This model performs interactive gap-filling to the temporal missing values by extrapolating data iteratively and has proven to obtain statistically consistent results (Rigling, 2012).

2.2.2. Support vector machine method

SVMs are based on statistical learning theory (Vapnik, 1995) and are commonly used for different applications, such as classification, regression estimation or pattern recognition (Camps-Valls et al., 2004; Gómez-Chova et al., 2010; Pal, 2006; Xie et al., 2008). This technique

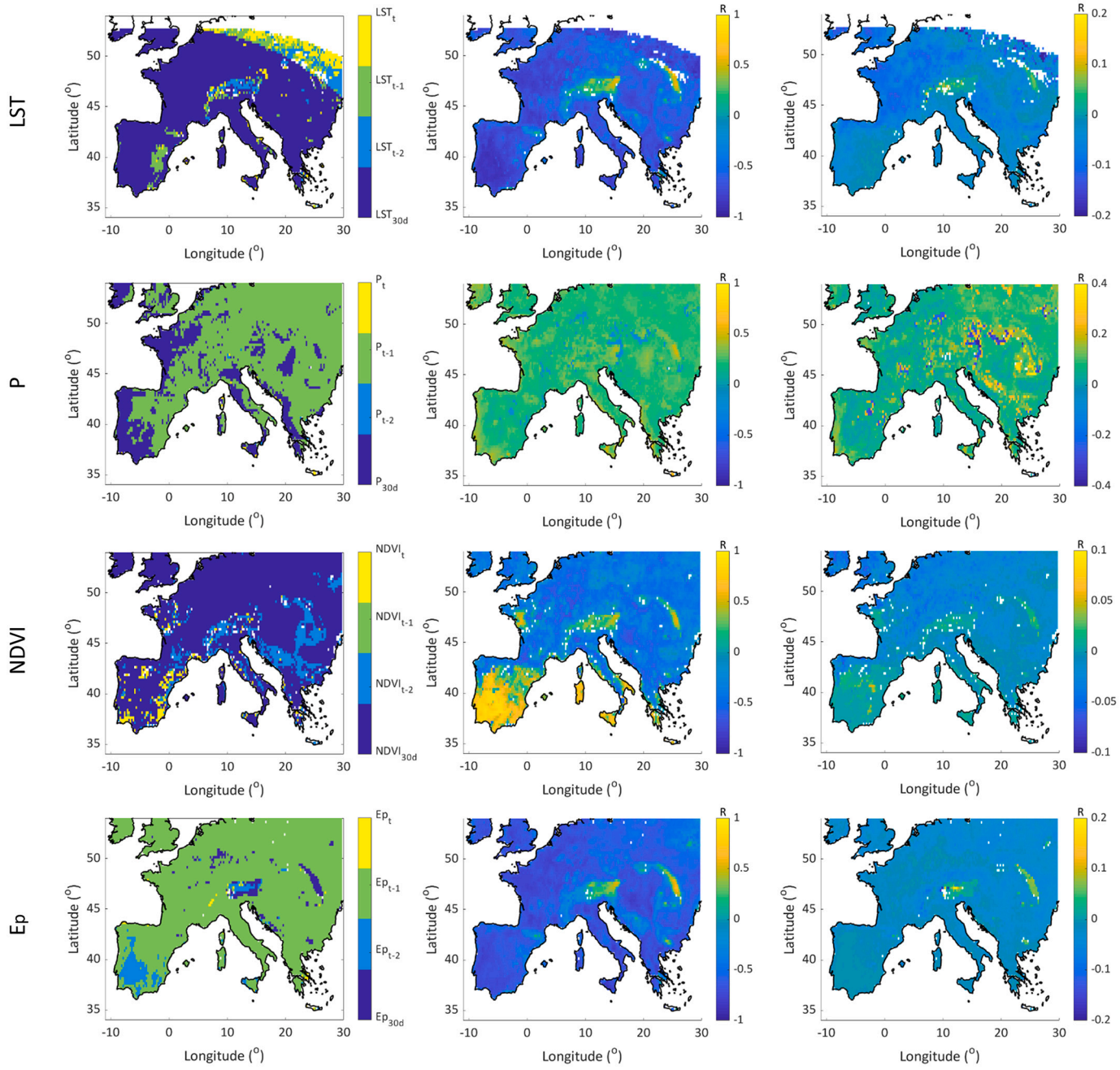

Fig. 3. Relationship between the ESA CCI SM and complementary variables LST, P, NDVI and E_p (from top to bottom). Maps on the left column show the type of series with the strongest correlation with SM for each pixel. Maps in the middle column show the value of the Pearson correlation coefficient (R) of the majoritarian type of series in the left maps. The maps on the right column show the difference between the maximum R and the second highest R.

was first developed for classification problems with the idea of obtaining the optimum separating hyperplane that maximizes the margin between patterns. Only the points that lie in the margin define the hyperplane, *i.e.*, the data points most difficult to classify (Dibike et al., 2001). Later, it was developed for regression problems, where the goal consists of finding a function $f(x)$ that describes the dependency between the inputs (x_i) and the target output $\boldsymbol{y}$ with at most an ε deviation and at the same time being as flat as possible (Smola and Schölkopf, 2004). Last, to consider nonlinear relationships, the kernel functions were added to the SVM algorithm. In this study, a Gaussian kernel function was used for that purpose and all the input variables were standardized before their incorporation to the models.

The SVM method has been extensively used in the remote sensing field due to its ability to generalize (Mountrakis et al., 2011), even with small training datasets. It has also proven to be a useful tool in SM estimation (Ahmad et al., 2010; Gill et al., 2006). Here, we implement the SVM model to estimate the missing values of the CCI SM series using all the variables of Table 1 as inputs. Khellouk et al. (2019) used these same variables to model SM with multiple linear regression methods. Nevertheless, not all the relationships between these variables and SM are linear, and sometimes disagreements exist between the changes in these variables and the SM response (Daly and Porporato, 2005). For this reason, a previous evaluation of these relationships was made. The level of correlation between the SM series and the complementary variables (X) series was studied in four different scenarios (the coincident series X_t, the series one day before X_{t-1}, the series two days before X_{t-2} and the series smoothed by averaging with a centered 30-day window X_{30d}). For each pixel, the correlation of SM with the different series of each variable was determined. The series with the highest correlation was identified, and the difference between the maximum correlation and the second highest correlation was calculated for each variable (Fig. 3). In view of the results, the series of LST_{30d}, P_{t-1}, $NDVI_{30d}$ and Ep_{t-1} were chosen since they provided the highest correlation for most of the pixels in the south of Europe. For NDVI, we decided to use the coincident series since it was already a smoothed series due to the daily interpolation of 16-day average values, and the differences obtained between the two maximum correlations were not remarkable.

In order to evaluate the ability of the SVM algorithm to estimate missing SM values, input variables were grouped. This approach allowed us to detect the crucial input variables for the SM estimation or those that could be expendable in the model. The groups were chosen based on the type of variables used; dynamic *d* (LST, P, NDVI and Ep), static *s* (soil texture and coordinates) or all of them *a* and on whether they used the SM_{t-1} series *h*. Thus, six groups were created for the spatial study *S* (Sdh, Sd, Ssh, Ss, Sah and Sa) and two were created for the temporal study *T*, where dynamic variables were used (Td and Tdh) (see Table 2).

A combination of SVM with PCA was also carried out. All input variables were considered for the analysis, *i.e.*, group Sah for spatial and group Tdh for temporal. The components explaining up to 95% of the variance were used as inputs to the SVM model.

Table 2
Groups of input variables used to estimate missing SM values with the SVM model in the spatial and temporal domains.

	SVM	Inputs
Spatial	Sah	SM_{t-1} + LST_{30d} + P_{t-1} + $NDVI_{30d}$ + Ep_{t-1} + Soil texture + Latitude + Longitude
	Sa	LST_{30d} + P_{t-1} + $NDVI_{30d}$ + Ep_{t-1} + Soil texture + Latitude + Longitude
	Ssh	SM_{t-1} + Soil texture + Latitude + Longitude
	Ss	Soil texture + Latitude + Longitude
	Sdh	SM_{t-1} + LST_{30d} + P_{t-1} + $NDVI_{30d}$ + Ep_{t-1}
	Sd	LST_{30d} + P_{t-1} + $NDVI_t$ + Ep_{t-1}
Temporal	Tdh	SM_{t-1} + LST_{30d} + P_{t-1} + $NDVI_{30d}$ + Ep_{t-1}
	Td	LST_{30d} + P_{t-1} + $NDVI_{30d}$ + Ep_{t-1}

2.2.3. Random forest method

The RF consists on a combination of independent tree predictors that depend on random vectors with the same distribution for all trees but that are sampled independently (Breiman, 2001). As the SVM, this method can be applied to both classification (Pal, 2005) and regression problems (Mutanga et al., 2012). In this study, the RF regression with 500 tree predictors was used to fill CCI SM gaps. The accuracy of the method was measured in the same way as for the other methods, thus the internal errors and correlation of the RF model were not considered. This ensured the results obtained with the different approaches are comparable. The combinations of input variables chosen for the spatial and temporal domain were the ones leading to the best estimates of SM using the SVMs.

3. Results and discussion

3.1. Temporal analysis

Six different approaches were evaluated (LI, CI, AR, SVM-Tdh, SVM-Td and SVM-PCA) to complete the CCI SM missing data in the temporal domain, *i.e.*, to complete the time series of each pixel individually. The obtained parameters in the cross-validation analysis (Fig. 4) show a good relationship between the estimated SM and the original SM CCI series for most of the pixels. The medians of R obtained in each approach range between 0.64 and 0.82 for CI and SVM-Tdh, respectively. The medians of the biases are practically zero in all cases. Some pixels provided negative (underestimation) and positive (overestimation) biases, but they very rarely exceeded 0.01 m^3m^{-3} in absolute values. There were hardly any differences between the RMSE and cRMSE due to the low bias; its medians range from 0.036 to 0.026 m^3m^{-3} for SVM with PCA and SVM-Tdh, respectively. These results are in line with those reported by Dumedah et al. (2014). The simplest methods resulted in the lowest accuracy, and the best results were obtained using a nonlinear autoregressive neural network, comparable to SVM-Tdh. When validation with the *in situ* SM series was performed (see Table 3), good accordance was obtained with SM CCI combined v4.5 product despite of the data gaps. Moreover, when reconstructed series were validated, the results obtained were also very similar to all the approaches. However, the correlation increases in most cases with the reconstructed CCI series by the SVM-Tdh, while the bias slightly increases or decreases, depending on the network. These results confirm that the SVM-Tdh method allows us to recreate the temporal dynamics of the original series without introducing significant errors.

While the obtained statistical scores seem to prove the effectiveness of all methods, a further examination of the results reveals that some of them might not be adequate. The LI, CI and AR methods have the advantage of being simple algorithms that require little processing time and do not need any auxiliary variables. However, we observed that their accuracy was compromised when dealing with large gaps (Fig. 5). Among ML-based algorithms, the SVM with the original variables (no PCA applied) could capture the temporal dynamics, while the SVM with PCA failed. These results suggest that PCA is not well suited for this problem, perhaps because all variables are interdependent or because the number of variables used in the PCA is not high enough and the resulting principal components introduce noise to the SVM models. The SVM without PCA worked best, the chosen variables had already shown their ability to capture the SM dynamics (Khellouk et al., 2019), and the SVM could reproduce it. A clear improvement is seen when the SM_{t-1} series is incorporated into the SVM, in agreement with previous studies (Gill et al., 2006). Thus, the SVM-Tdh has proven to be the most efficient and accurate approach. The R obtained is higher than 0.7 in 85% of the study area, similar to that obtained by Cui et al. (2019) and Wang et al. (2012). An analysis of the spatial distributions of R showed that the poorest values were located in mountainous regions and highest values of RMSE were obtained in coastal areas and in the Balkan Peninsula (Fig. 6). We observed that values of R lower than 0.6 corresponded to

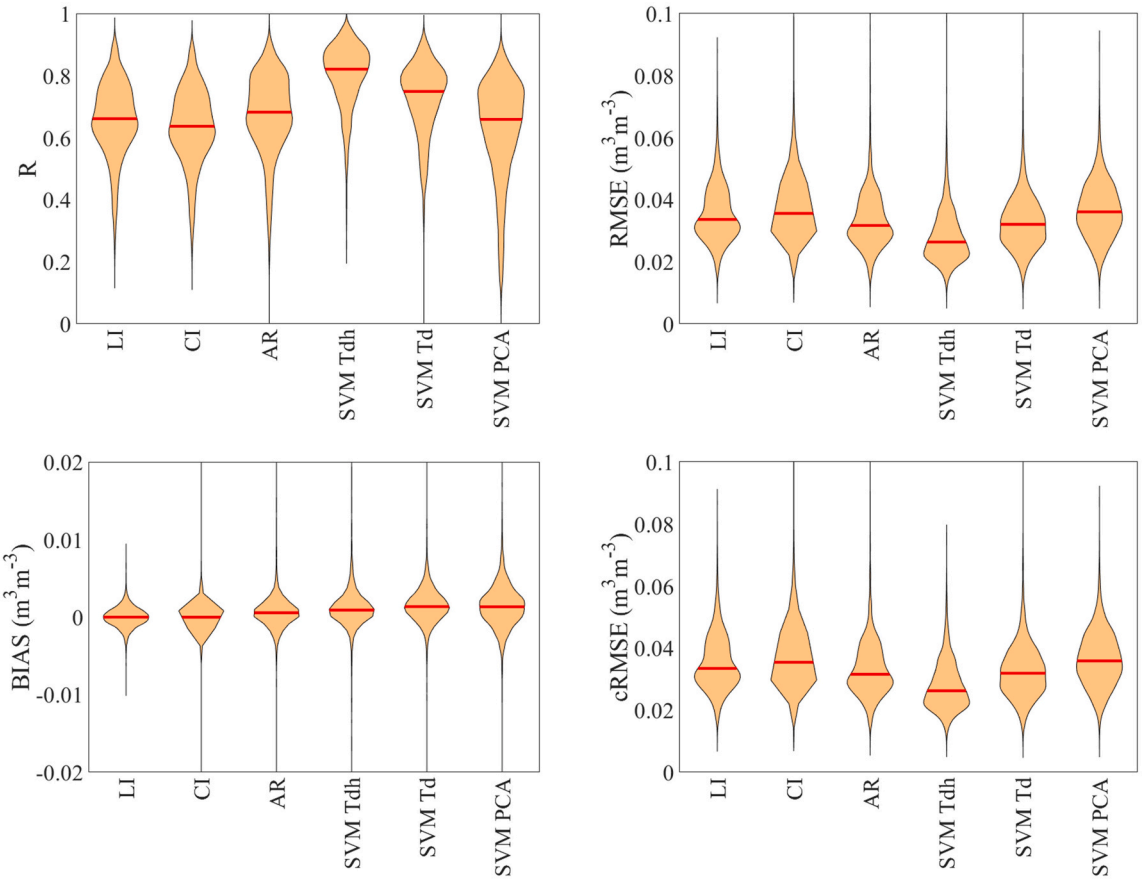

Fig. 4. Statistical parameters of the cross-validation obtained for the six approaches of the CCI SM gap-filling procedure in the temporal domain.

Table 3
Validation of the CCI SM series, original and reconstructed in the temporal domain, with *in situ* networks.

CCI SM	R				BIAS (m^3m^{-3})			
	ORACLE	TERENO	REMEDHUS	UMBRIA	ORACLE	TERENO	REMEDHUS	UMBRIA
Original	0.51	0.63	0.83	0.63	−0.014	0.019	−0.099	−0.004
LI	0.50	0.60	0.84	0.66	−0.015	0.019	−0.100	0.001
CI	0.50	0.59	0.84	0.66	−0.015	0.019	−0.100	0.001
AR	0.50	0.62	0.84	0.68	−0.015	0.018	−0.101	0.001
SVM Tdh	0.49	0.65	0.84	0.71	−0.016	0.018	−0.100	−0.001
SVM Td	0.48	0.64	0.83	0.70	−0.016	0.018	−0.101	−0.001
SVM PCA	0.50	0.65	0.82	0.67	−0.014	0.018	−0.101	−0.001

pixels with noisy SM series, *i.e.*, with non-seasonal behaviour or with a high uncertainty of the original CCI SM values. This result suggests that the gap-filling performance depends on the quality of the original SM values to some extent. In fact, the mean CCI SM uncertainty and the RMSE obtained for each pixel showed a correlation coefficient of 0.78. A negative relationship between the correlation coefficient of the gap-filling method and the percentage of available data also exists but is lower (R = -0.64). This result could be attributed directly to the introduction of the SM series in the algorithm. However, this was also observed with all the methods. Ahmad et al. (2010) obtained lower accuracies at stations with a high vegetation density using SVMs. In this study no significant relationship was found between the accuracy and the mean values of NDVI, but a relationship was found between the RMSE obtained with all the methods and E_p mean values (R ranging between 0.51 and 0.64).

The SVM-Tdh has been proven to be accurate enough to complete the CCI SM series gaps for the study period. However, the availability of complementary databases with a coincident period is probably a limitation to complete the 40-year series. The SVMs do not determine the influence of each input variable on the output variable. However, with the aim of exploring whether some of the input variables would be expendable or crucial, all possible combinations were performed, starting with all of them and progressively eliminating one at a time until only the SM_{t-1} remained, similar to what Yang et al. (2006) did. A total of 16 combinations were tested, and they were separated into groups based on the number of variables, which ranged from 5 to 1. The best results were obtained with the five variables (Table 4), *i.e.*, with the original SMV-Tdh approach, but hardly any differences were observed with the remaining combinations. The accuracy was slightly reduced as the number of variables decreased.

Moreover, it was observed that the variables of the best datasets were consistent, showing more importance to LST, followed by P, NDVI and E_p (not shown). These results imply that it is possible to obtain good performance with the SVM model even if some of the input variables are not available. As expected, using only SM is not enough, and the results obtained in this case are equivalent to those obtained with simple

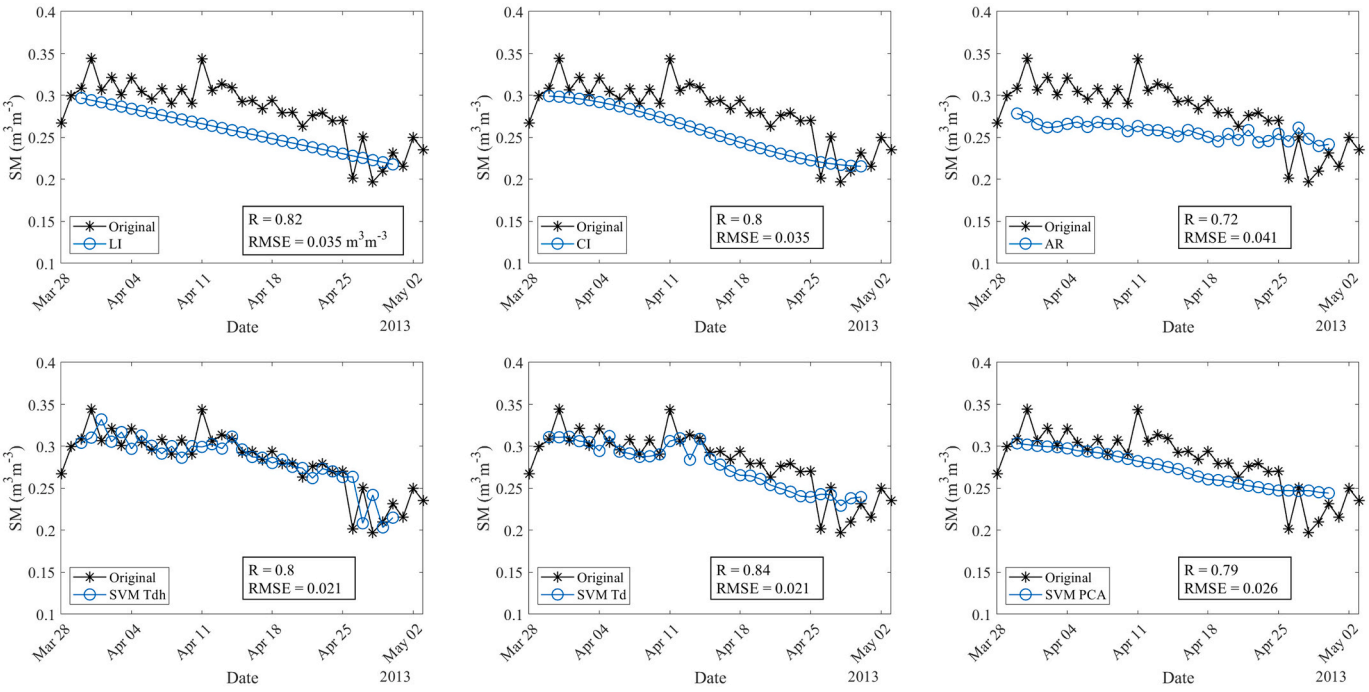

Fig. 5. Pixel SM series reconstruction for each temporal approach. The original CCI SM series (black) and the reconstructed series according to a test set (blue) are shown for a period of three months. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

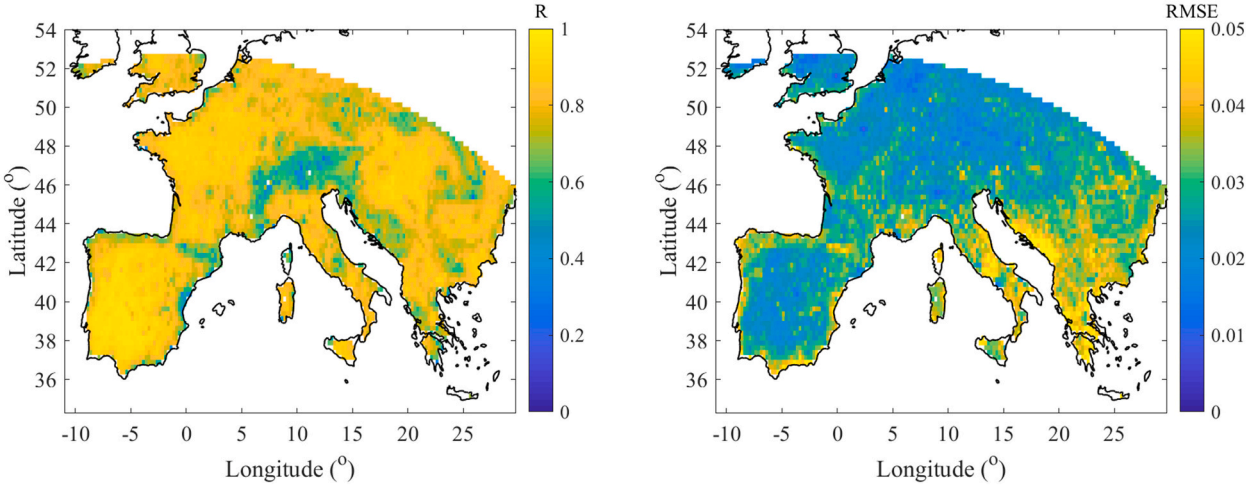

Fig. 6. Spatial distribution of R (right) and RMSE (left) obtained with SVM Tdh in the temporal series study.

Table 4
Median values of the best combination for each group for the input variable combination analysis in the temporal domain.

N° Variables	Combination	R	bias (m^3m^{-3})	RMSE (m^3m^{-3})	cRMSE (m^3m^{-3})
5	SM_{t-1}, LST_{30d}, P, $NDVI_{30d}$, E_p	0.834	0.000	0.025	0.025
4	SM_{t-1}, LST_{30d}, P, $NDVI_{30d}$	0.831	0.000	0.026	0.025
3	SM_{t-1}, LST_{30d}, P	0.826	0.000	0.026	0.026
2	SM_{t-1}, LST_{30d}	0.798	0.000	0.026	0.026
1	SM	0.755	−0.001	0.028	0.028

interpolations.

3.2. Spatial analysis

For the spatial domain, missing values were estimated by applying a total of nine approaches to each CCI SM daily image. These methods were LI, CI, SVM-Sah, SVM-Sa, SVM-Ssh, SVM-Ss, SVM Sdh, SVM-Sd and SVM with PCA (Table 2). The results of the cross-validation analysis show that most methods yield adequate SM estimations (Fig. 7). The median values of R ranged between 0.48 and 0.88 for SVM with PCA and SVM-Ssh, respectively. The bias values, as for the temporal domain, are practically zero in all cases, implying few differences between the RMSE and cRMSE values. The RMSE and cRMSE medians range between 0.043 and 0.024 m^3m^{-3} for SVM with PCA and SVM-Ssh, respectively. These results are similar to those obtained by Llamas et al. (2020), who reported higher RMSE values than those obtained with SVM-Ssh and

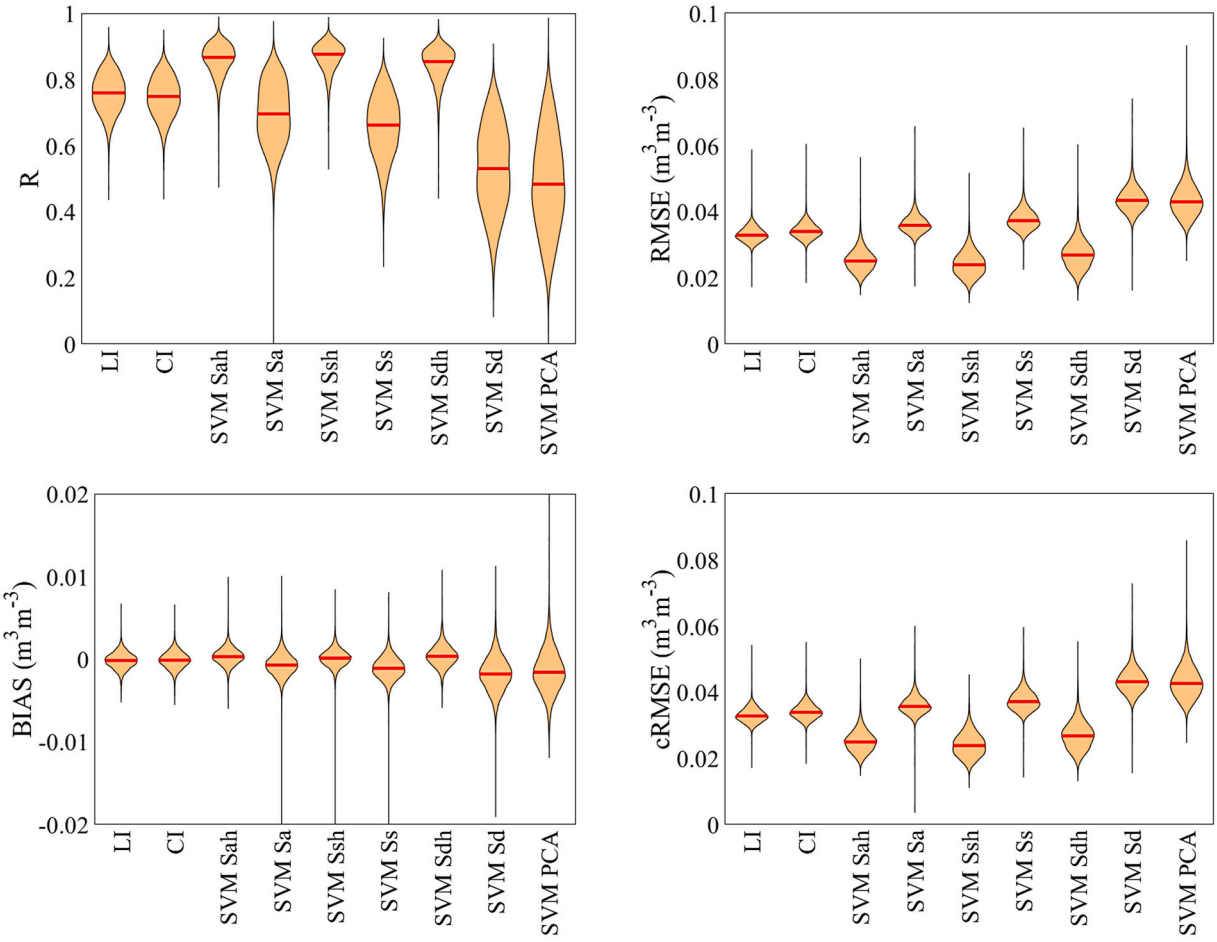

Fig. 7. Statistical parameters of the cross-validation obtained for the nine approaches of the CCI SM gap-filling procedure in the spatial domain.

similar correlation coefficients with the ordinary kriging and regression kriging methods. In addition, their results with the generalized linear models showed lower R and higher RMSE than those obtained with LI and CU. Although our results cannot be directly comparable as they were focused on different study regions, they nonetheless provide an idea about the quality of each method to estimate the missing CCI SM values. The results obtained in the validation with *in situ* data (see Table 5) showed lower correlations between the reconstructed series and the ground measurements when using any of the interpolation methods. However, SVMs without using SM_{t-1} series offer similar results as the interpolations approaches (Fig. 7). When the SM_{t-1} is used, the reconstructed series validation results barely differ from the ones obtained with the original series, confirming a good ability to estimate the CCI SM missing values in this input data configuration.

Again, it was observed that the SVM method performed better than the other methods but only when SM_{t-1} was used. SVM with PCA, as was found with the temporal domain, showed the poorest SM estimations (Fig. 8). In addition, similar results were obtained with SVM-Sd (*i.e.*, dynamic variables as inputs and without the SM_{t-1} series). Hence, it follows that the SVM method is not capable of accurately simulating the spatial dynamics of SM without previous information on SM. Nevertheless, when static variables are added to the SVM (Sah, Ssh), the results improve.

These variables describe the soil properties and place SM values in space. SVM-Ssh obtained slightly better results than did SVM-Sah. Although differences in accuracy are small, SVM-Ssh has the advantage of using the lowest number of variables. Furthermore, the constraint of the availability of auxiliary databases with a coincident

Table 5
Validation of the CCI SM series, original and reconstructed in the spatial domain, with *in situ* networks.

CCI SM	R				BIAS (m^3m^{-3})			
	ORACLE	TERENO	REMEDHUS	UMBRIA	ORACLE	TERENO	REMEDHUS	UMBRIA
Original	0.51	0.63	0.83	0.63	−0.014	0.019	−0.099	−0.004
LI	0.50	0.56	0.80	0.63	−0.015	0.021	−0.095	−0.003
CI	0.49	0.55	0.79	0.62	−0.015	0.021	−0.095	−0.002
SVM Sah	0.51	0.64	0.83	0.66	−0.012	0.019	−0.099	−0.001
SVM Sa	0.50	0.63	0.82	0.68	−0.014	0.020	−0.098	−0.008
SVM Ssh	0.51	0.64	0.83	0.66	−0.012	0.019	−0.099	−0.001
SVM Ss	0.51	0.62	0.82	0.67	−0.014	0.021	−0.098	−0.011
SVM Sdh	0.51	0.64	0.83	0.66	−0.012	0.019	−0.099	0.000
SVM Sd	0.50	0.64	0.82	0.70	−0.014	0.020	−0.098	−0.005
SVM PCA	0.51	0.65	0.81	0.63	−0.014	0.017	−0.096	−0.007

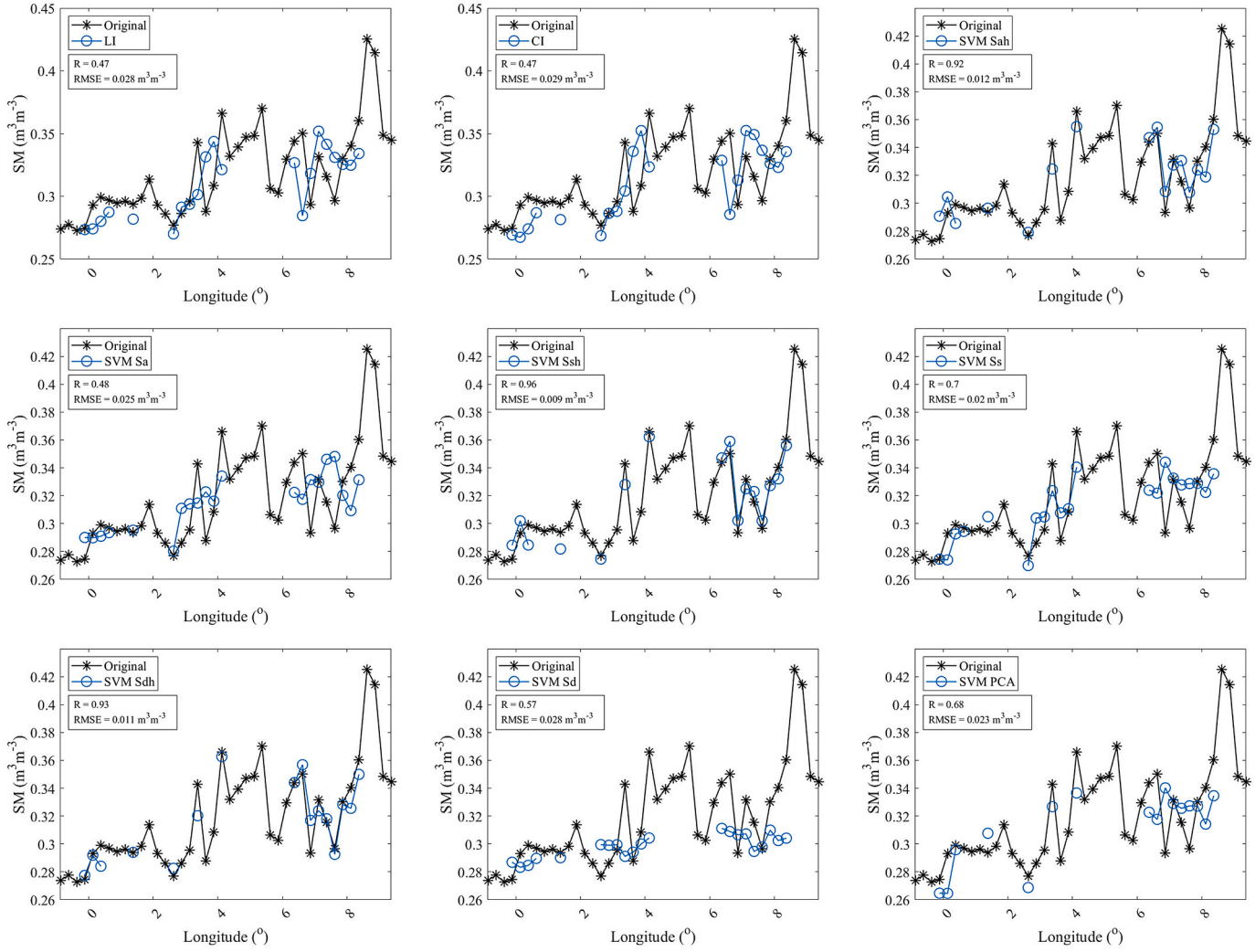

Fig. 8. A transect for latitude 47.875° of the November 10, 2012, SM series reconstruction for each spatial approach. The original CCI SM series (black) and the reconstructed series according to a test set (blue). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

period with that of the CCI SM series disappears. To our knowledge, this is the first study that applies ML techniques in the spatial domain to estimate SM; therefore, we cannot compare our obtained results with those obtained using similar approaches. It should be noted that although the vast majority of gap-filling studies do not consider the use of only static layers, our results show that this could be a promising rapid and good approach to address this problem.

An analysis for assessing the relevance of the input variables was also carried out for SVM-Ssh, the best performing approach in the spatial domain. Six variables were combined and evaluated in groups, as was done for the temporal domain. In total, 32 combinations in 6 groups were studied, always using the SM_{t-1}. The best result obtained for each group was very similar in all cases (Table 6). Unlike the temporal study, the most accurate estimations were not obtained with the original combination of variables, that is, with the six variables, but with using only the geographic coordinates of each pixel and the sand content. This result implies that the SVM method could be simplified by reducing the number of input variables and reach comparable performances.

Table 6
Median values of the best combination for each group for the input variable combination analysis in the spatial domain.

N° Variables	Combination	R	bias (m^3m^{-3})	RMSE (m^3m^{-3})	cRMSE (m^3m^{-3})
6	SM_{t-1}, sand, clay, silt, lat., lon.	0.875	0.000	0.024	0.024
5	SM_{t-1}, sand, silt, lat., lon.	0.878	0.000	0.024	0.024
4	SM_{t-1}, sand, lat., lon.	0.879	0.000	0.024	0.024
3	SM_{t-1}, lat., lon.	0.869	0.000	0.025	0.025
2	SM_{t-1}, lat.,	0.836	0.000	0.027	0.027
1	SM_{t-1}	0.818	0.000	0.029	0.029

3.3. Random forest with the best combination of input variables

The selection of the input variables to the SVM has been key to obtain good estimates of SM. To test the capacity of other non-linear approaches to estimating CCI SM missing values, RF was evaluated using the best combination of input variables obtained with SVMs for spatial and temporal domains. Thus, for the temporal domain, the variables of the SVM Tdh approach were used and for the spatial, the coordinates and the sand content together with the SM_{t-1} series were used. The results (Fig. 9) show that for the temporal domain the correlation is similar to that obtained with the SVM Tdh, while the errors are slightly higher with a median of RMSE = 0.029 m^3m^{-3}. RF has proven to be a good tool when estimating SM series in other studies (Long et al., 2019; Zhao et al.,

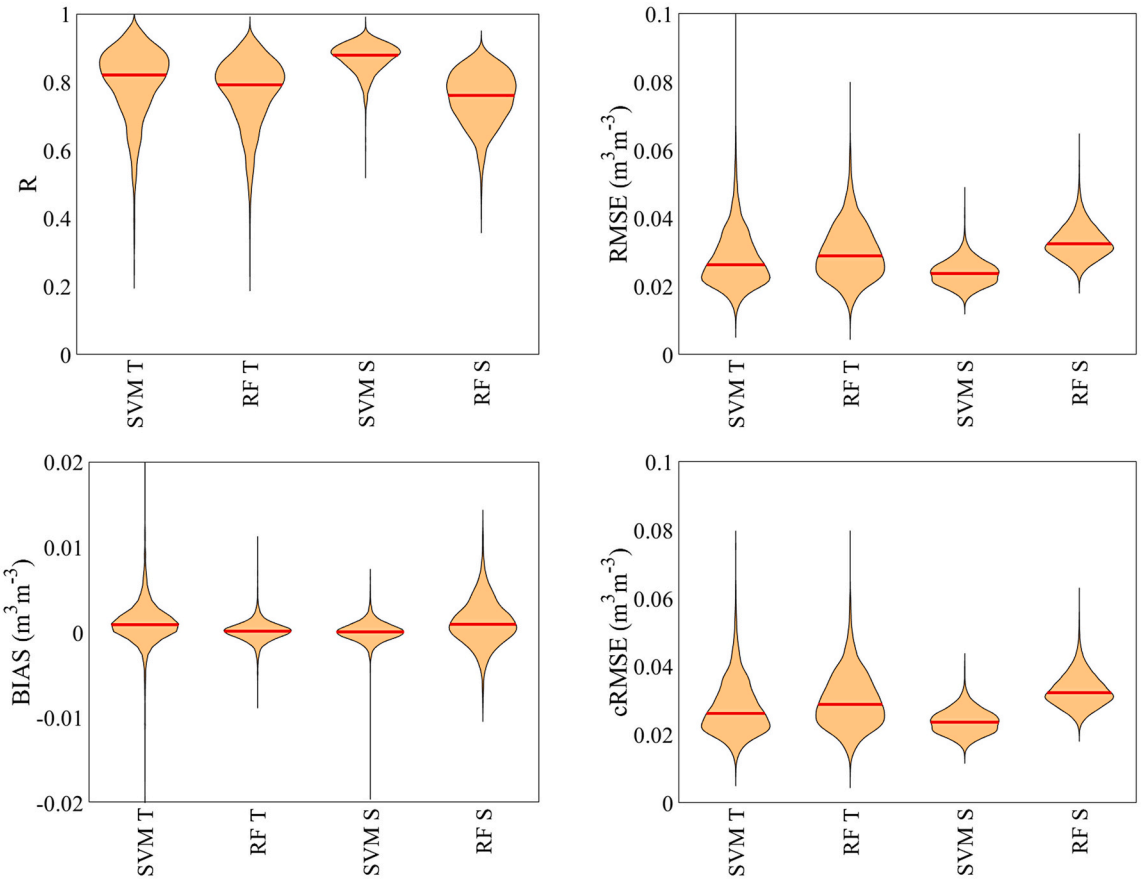

Fig. 9. Statistical parameters of the cross-validation obtained for RF and SVM with best combination of input variables in the spatial (S) and temporal (T) domain.

2018) and results obtained here are in line with them. However, the SVM showed to be more accurate. On the contrary, the RF in spatial domain do not seem to be as accurate as SVM. The median of R obtained with RF is 0.76, while with SVM is 0.88. The bias and the errors obtained with RF are also higher, which indicates that RF in the spatial domain is not able to generalize as well as the SVM.

3.4. Comparison of the most accurate spatial and temporal approaches

In general, all the methods and the different approaches studied in this work have proven to be valid gap-filling techniques for the CCI SM series. However, some of them reconstructed the time series with greater precision, specifically the SVM methods. For the temporal domain study, the SVM using all the dynamic auxiliary variables and the SM_{t-1} as inputs provided the best results. For the spatial domain study, the SVM with only the coordinates, sand content and SM of the day before as inputs was the most efficient. The percentiles of the parameters obtained in the cross-validation analysis were calculated with the aim of comparing the performances of the two approaches (Fig. 10). The correlation was greater for the spatial approach in all of the cases, and the RMSE was slightly lower. In addition, we observed that the temporal approach presented more outliers than the spatial approach. This result is possibly due its relationship with the uncertainty of the original CCI SM value in the temporal approach, while the spatial approach unifies all the pixels and, consequently, leads to more stable estimates. Despite being an uncommon approach, in terms of accuracy, the SVM applied in the spatial domain seems to be more appropriate to reconstruct CCI SM series.

Regarding the applicability of the methods, on the one hand, the temporal approach enables the completion of the series of each pixel, while the spatial approach depends on the availability of the SM value of the day before. Hence, it would be necessary to iteratively calculate the SM as is done in the temporal approach, but the quantification of the accuracy of each model would not be plausible since many SVM models would be mixed (one for each day). On the other hand, the spatial approach has the advantage of using only static auxiliary variables and could be used seamlessly to complete the 40 years of the CCI SM series. However, the precision of the method likely decreases in the early years of the series, where data availability is very low. This should be the subject of further research. Additionally, the spatial approach cannot be conducted on days where no SM value is available in an entire region, unlike the temporal approach where there will always be values for a given pixel. In summary, the performance of the two kinds of approaches may be linked to the data availability for a specific region and period, but the two are complementary and show great potential in terms of estimating the missing values in the CCI SM database.

4. Conclusions

The CCI SM database is currently the longest available data record of satellite soil moisture. However, its applicability is often compromised due to its spatiotemporal gaps, which depend mostly on the available satellites across its 40-year period. In this paper, gap-filling techniques of different complexities have been applied to ESA CCI SM data over Southern Europe for the period 2003–2015. Methods focused on the spatial and temporal domains have been explored and compared. Our results show that interpolation methods in both the temporal and the spatial domains have the advantage of being simple. However, when dealing with large spatial or temporal gaps, they failed. The autoregressive method in the temporal study somewhat improved the estimates obtained by the interpolations but could not address the large data gaps. In contrast, SVMs have proven to be a very robust technique

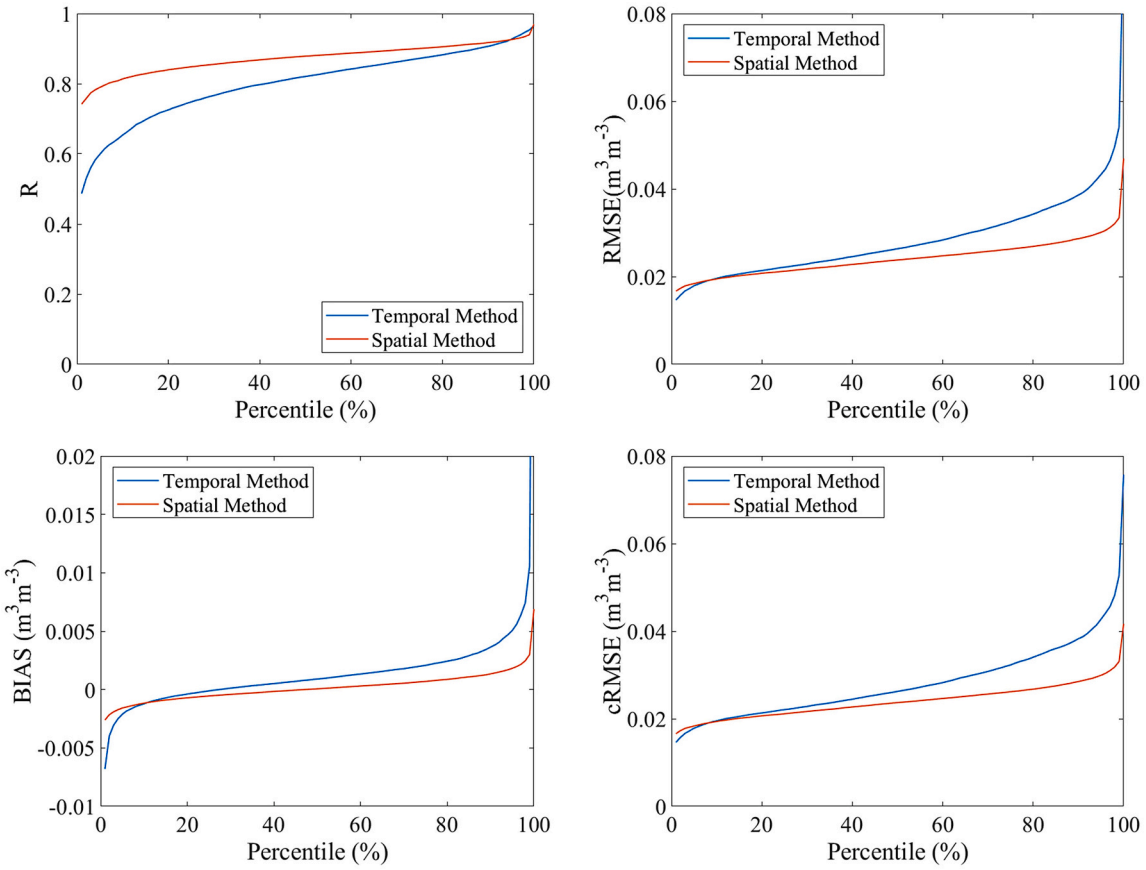

Fig. 10. Percentiles of the cross-validation analysis parameters for the most accurate temporal (blue) and spatial (orange) approaches evaluated in this work. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

for estimating missing values, especially in the spatial domain, and the accuracy of the estimates hardly depends on the gap length. The effectiveness of the SVMs had previously been tested on *in situ* SM databases but never on satellite databases. The choice of input variables for the SVM has been shown to have an impact on the method's performance. Using the SM_{t-1} series has proven to be key to obtaining good estimations in both temporal and spatial analysis. The application of PCA to input variables did not improve the performance and actually led to significantly worse results. In addition, the RF were not able to reproduce the spatial dynamics of the SM as well as the SVM did, and the obtained precision was slightly lower than the SVM when applied in the temporal domain.

For the temporal study, it was observed that the chosen variables were able to reproduce the SM dynamics. In particular, LST and precipitation proved to be a crucial variable since the accuracy always decreased when it was not incorporated. This result suggests that SVMs are able to capture the nonlinear relationship between the dynamics of the two variables. Interestingly, the variables used in the temporal approaches did not prove to be as useful in the spatial domain, where the best results were obtained with variables describing the spatial distribution of soil physical properties and antecedent SM. When evaluating the relevance of input variables, we obtained the best results when only the coordinates together with sand content and the SM one day before were incorporated into the SVM. This approach presents an advantage since the number of input variables is reduced, and therefore, the SVM model is simplified. However, it has been shown that the two approaches are complementary, so a combination of them could resolve some of their limitations. This assumption might be addressed in future studies.

The applicability of all the methods has been validated, and the results are satisfactory. We proved that, despite being an uncommon approach, the SVMs in the spatial domain, with the appropriate input variables, can be a suitable method used to fill the gaps in the CCI SM database. Thus, this approach can be used to obtain a long-term SM satellite-derived series with homogeneous coverage, meaning there are fewer limitations in applications where data gaps are a problem. Overall, our results demonstrate the applicability of ML approaches for gap-filling multidecadal soil moisture observational data records and highlight the value of exploiting both the temporal and the spatial domains.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study was supported by the Spanish Ministry of Science, Innovation and Universities (Projects ESP2017-89463-C3-3-R and RTI2018-096765-A-100), the Castilla y León Government (Project SA112P20) and the European Regional Development Fund (ERDF). The authors also acknowledge the project Unidad de Excelencia CLU-2018-04 co-funded by ERDF and Castilla y León Government.

References

Ahmad, S., Kalra, A., Stephen, H., 2010. Estimating soil moisture using remote sensing data: a machine learning approach. Adv. Water Resour. 33, 69–80. https://doi.org/10.1016/j.advwatres.2009.10.008.

Ali, I., Greifeneder, F., Stamenkovic, J., Neumann, M., Notarnicola, C., 2015. Review of machine learning approaches for biomass and soil moisture retrievals from remote sensing data. Remote Sens. 7, 16398–16421. https://doi.org/10.3390/rs71215841.

Al-Yaari, A., Wigneron, J.P., Dorigo, W., Colliander, A., Pellarin, T., Hahn, S., Mialon, A., Richaume, P., Fernandez-Moran, R., Fan, L., Kerr, Y.H., De Lannoy, G., 2019. Assessment and inter-comparison of recently developed/reprocessed microwave satellite soil moisture products using ISMN ground-based measurements. Remote Sens. Environ. 224, 289–303. https://doi.org/10.1016/j.rse.2019.02.008.

An, R., Zhang, L., Wang, Z., Quaye-Ballard, J.A., You, J., Shen, X., Gao, W., Huang, L., Zhao, Y., Ke, Z., 2016. Validation of the ESA CCI soil moisture product in China. Int. J. Appl. Earth Obs. and Geoinf. 48, 28–36. https://doi.org/10.1016/j.jag.2015.09.009.

Ballabio, C., Panagos, P., Monatanarella, L., 2016. Mapping topsoil physical properties at European scale using the LUCAS database. Geoderma 261, 110–123. https://doi.org/10.1016/j.geoderma.2015.07.006.

Beck, H.E., Pan, M., Miralles, D.G., Reichle, R.H., Dorigo, W.A., Hahn, S., Sheffield, J., Karthikeyan, L., Balsamo, G., Parinussa, R.M., van Dijk, A.I.J.M., Du, J., Kimball, J. S., Vergopolan, N., Wood, E.F., 2020. Evaluation of 18 satellite- and model-based soil moisture products using *in situ* measurements from 826 sensors. Hydrol. Earth Syst. Sci. Discuss. https://doi.org/10.5194/hess-2020-184 (in review).

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Brocca, L., Melone, F., Moramarco, T., 2008. On the estimation of antecedent wetness conditions in rainfall–runoff modelling. Hydrol. Process. 22, 629–642. https://doi.org/10.1002/hyp.6629.

Brocca, L., Melone, F., Moramarco, T., 2011. Distributed rainfall-runoff modelling for flood frequency estimation and flood forecasting. Hydrol. Process. 25, 2801–2813. https://doi.org/10.1002/hyp.8042.

Browne, M.W., 2000. Cross-validation methods. J. Math. Psychol. 44, 108–132. https://doi.org/10.1006/jmps.1999.1279.

Camps-Valls, G., Gómez-Chova, L., Calpe-Maravilla, J., Martín-Guerrero, J.D., Soria-Olivas, E., Alonso-Chordá, L., Moreno, J., 2004. Robust support vector method for hyperspectral data classification and knowledge discovery. IEEE Trans. Geosci. Remote Sens. 42, 1530–1542. https://doi.org/10.1109/TGRS.2004.827262.

Champagne, C., White, J., Berg, A., Belair, S., Carrera, M., 2019. Impact of soil moisture data characteristics on the sensitivity to crop yields under drought and excess moisture conditions. Remote Sens. 11, 372. https://doi.org/10.3390/rs11040372.

Ciabatta, L., Massari, C., Brocca, L., Gruber, A., Reimer, C., Hahn, S., Paulik, C., Dorigo, W., Kidd, R., Wagner, W., 2018. SM2RAIN-CCI: a new global long-term rainfall data set derived from ESA CCI soil moisture. Earth Syst. Sci. Data Discuss 10, 267–280. https://doi.org/10.5194/essd-2017-86.

Copernicus Climate Service (C3S), 2020. https://climate.copernicus.eu (accessed 22 June 2020).

Cui, Y., Long, D., Hong, Y., Zeng, C., Zhou, J., Han, Z., Liu, R., Wan, W., 2016. Validation and reconstruction of FY-3B/MWRI soil moisture using an artificial neural network based on reconstructed MODIS optical products over the Tibetan plateau. J. Hydrol. 543, 242–254. https://doi.org/10.1016/j.jhydrol.2016.10.005.

Cui, Y., Zeng, C., Zhou, J., Xie, H., Wan, W., Hu, L., Xiong, W., Chen, X., Fan, W., Hong, Y., 2019. A spatio-temporal continuous soil moisture dataset over the Tibet plateau from 2002 to 2015. Sci. data 6, 247. https://doi.org/10.1038/s41597-019-0228-x.

Daly, E., Porporato, A., 2005. A review of soil moisture dynamics: from rainfall infiltration to ecosystem response. Environ. Eng. Sci. 22, 9–24. https://doi.org/10.1089/ees.2005.22.9.

Dibike, Y.B., Velickov, S., Solomatine, D., Abbott, M.B., 2001. Model induction with support vector machines: introduction and applications. J. Comput. Civ. Eng. 15 (3), 208–216. https://doi.org/10.1061/(ASCE)0887-3801(2001)15:3(208).

Didan, K., 2015. MOD13A2 MODIS/terra vegetation indices 16-day L3 global 1km SIN grid V006. In: NASA EOSDIS Land Processes DAAC. https://doi.org/10.5067/MODIS/MOD13A2.006.

Dorigo, W., Bauer-marschallinger, B., Depoorter, M., Miralles, D., 2016. Assessing the Impact of the 2015 / 2016 El Niño Event on Multi-Satellite Soil Moisture over the Southern, 18. Hemisphere, p. 15476.

Dorigo, W., Wagner, W., Albergel, C., Albrecht, F., Balsamo, G., Brocca, L., Chung, D., Ertl, M., Forkel, M., Gruber, A., Haas, E., Hamer, P.D., Hirschi, M., Ikonen, J., de Jeu, R., Kidd, R., Lahoz, W., Liu, Y.Y., Miralles, D., Mistelbauer, T., Nicolai-Shaw, N., Parinussa, R., Pratola, C., Reimer, C., van der Schalie, R., Seneviratne, S.I., Smolander, T., Lecomte, P., 2017. ESA CCI soil moisture for improved earth system understanding: state-of-the art and future directions. Remote Sens. Environ. 203, 185–215. https://doi.org/10.1016/j.rse.2017.07.001.

Dorigo, W.A., Wagner, W., Hohensinn, R., Hahn, S., Paulik, C., Xaver, A., Gruber, A., Drusch, M., Mecklenburg, S., Van Oevelen, P., Robock, A., Jackson, T., 2011. The international soil moisture network: a data hosting facility for global *in situ* soil moisture measurements. Hydrol. Earth Syst. Sci. 15, 1675–1698. https://doi.org/10.5194/hess-15-1675-2011.

Dorigo, W.A., Gruber, A., De Jeu, R.A.M., Wagner, W., Stacke, T., Loew, A., Albergel, C., Brocca, L., Chung, D., Parinussa, R.M., Kidda, R., 2015. Evaluation of the ESA CCI soil moisture product using ground-based observations. Remote Sens. Environ. 162, 380–395. https://doi.org/10.1016/j.rse.2014.07.023.

Duguay-Tetzlaff, A., Stöckli, R., Bojanowski, J., Hollmann, R., Fuchs, P., Werscheck, M., 2017. CM SAF land SUrface temperature dataset from METeosat first and second generation - edition 1 (SUMET Ed. 1). In: Satellite Application Facility on Climate Monitoring. https://doi.org/10.5676/EUM_SAF_CM/LST_METEOSAT/V001.

Dumedah, G., Coulibaly, P., 2011. Evaluation of statistical methods for infilling missing values in high-resolution soil moisture data. J. Hydrol. 400, 95–102. https://doi.org/10.1016/j.jhydrol.2011.01.028.

Dumedah, G., Walker, J.P., Chik, L., 2014. Assessing artificial neural networks and statistical methods for infilling missing soil moisture records. J. Hydrol. 515, 330–344. https://doi.org/10.1016/j.jhydrol.2014.04.068.

Entekhabi, D., Reichle, R.H., Koster, R.D., Crow, W.T., 2010. Performance metrics for soil moisture retrievals and application requirements. J. Hydrometeorol. 11, 832–840. https://doi.org/10.1175/2010JHM1223.1.

Feng, H., 2016. Individual contributions of climate and vegetation change to soil moisture trends across multiple spatial scales. Sci. Rep. 6, 1–6. https://doi.org/10.1038/srep32782.

Ford, T.W., Quiring, S.M., 2014. Comparison and application of multiple methods for temporal interpolation of daily soil moisture. Int. J. Climatol. 34, 2604–2621. https://doi.org/10.1002/joc.3862.

Fritsch, F.N., Carlson, R.E., 1980. Monotone piecewise cubic interpolation. SIAM J. Numer. Anal. 17, 238–246.

García, D., 2010. Robust smoothing of gridded data in one and higher dimensions with missing values. Comput. Stat. Data Anal. 54, 1167–1178. https://doi.org/10.1016/j.csda.2009.09.020.

GCOS, 2010. Implementation Plan for the Global Observing System for Climate (GCOS) in Support of the United Nations Framework Convention on Climate Change (UNFCCC). World Meteorological Organization, Geneva, Switzerland.

Gill, M.K., Asefa, T., Kemblowski, M.W., McKee, M., 2006. Soil moisture prediction using support vector machines. J. Am. Water Resour. Assoc. 42, 1033–1046. https://doi.org/10.1111/j.1752-1688.2006.tb04512.x.

Gómez-Chova, L., Camps-Valls, G., Bruzzone, L., Calpe-Maravilla, J., 2010. Mean map kernel methods for semisupervised cloud classification. IEEE Trans. Geosci. Remote Sens. 48, 207–220. https://doi.org/10.1109/TGRS.2009.2026425.

González-Zamora, Á., Sánchez, N., Martínez-Fernández, J., Wagner, W., 2016. Root-zone plant available water estimation using the SMOS-derived soil water index. Adv. Water Resour. 96, 339–353. https://doi.org/10.1016/j.advwatres.2016.08.001.

González-Zamora, Á., Sánchez, N., Pablos, M., Martínez-Fernández, J., 2019. CCI soil moisture assessment with SMOS soil moisture and *in situ* data under different environmental conditions and spatial scales in Spain. Remote Sens. Environ. 225, 469–482. https://doi.org/10.1016/j.rse.2018.02.010.

Gruber, A., Dorigo, W.A., Crow, W., Wagner, W., 2017. Triple collocation-based merging of satellite soil moisture retrievals. IEEE Trans. Geosci. Remote Sens. 55, 6780–6792. https://doi.org/10.1109/TGRS.2017.2734070.

Gruber, A., Scanlon, T., Van Der Schalie, R., Wagner, W., Dorigo, W., 2019. Evolution of the ESA CCI soil moisture climate data records and their underlying merging methodology. Earth Syst. Sci. Data 11, 717–739. https://doi.org/10.5194/essd-11-717-2019.

Huffman, G.J., Bolvin, D.T., Braithwaite, D., Hsu, K., Joyce, R., Xie, P., Yoo, S.H., 2019. NASA global precipitation measurement (GPM) integrated multi-satellite retrievals for GPM (IMERG). In: Algorithm Theoretical Basis Document (ATBD) Version, 6.

Ikonen, J., Smolander, T., Rautiainen, K., Cohen, J., Lemmetyinen, J., Salminen, M., Pulliainen, J., 2018. Spatially distributed evaluation of ESA CCI soil moisture products in a northern boreal Forest environment. Geosciences 8, 51. https://doi.org/10.3390/geosciences8020051.

Jackson, T.J., 1993. III. Measuring surface soil moisture using passive microwave remote sensing. Hydrol. Process. 7, 139–152. https://doi.org/10.1002/hyp.3360070205.

Khellouk, R., Barakat, A., El Jazouli, A., Boudhar, A., Lionboui, H., Rais, J., Benabdelouahab, T., 2019. An integrated methodology for surface soil moisture estimating using remote sensing data approach. Geocarto Int. https://doi.org/10.1080/10106049.2019.1655797.

Kornelsen, K., Coulibaly, P., 2014. Comparison of interpolation, statistical, and data-driven methods for imputation of missing values in a distributed soil moisture dataset. J. Hydrol. Eng. 19, 26–43. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000767.

Korres, W., Reichenau, T.G., Schneider, K., 2013. Patterns and scaling properties of surface soil moisture in an agricultural landscape: An ecohydrological modeling study. J. Hydrol. 498, 89–102. https://doi.org/10.1016/j.jhydrol.2013.05.050.

Lary, D.J., Alavi, A.H., Gandomi, A.H., Walker, A.L., 2016. Machine learning in geosciences and remote sensing. Geosci. Front. 7, 3–10. https://doi.org/10.1016/j.gsf.2015.07.003.

Liu, Y., Yang, Y., Jing, W., 2020. Potential applicability of SMAP in ECV soil moisture gap-filling: a case study in Europe. IEEE Access 8, 133114–133127. https://doi.org/10.1109/ACCESS.2020.3009977.

Liu, Yongwei, Liu, Yuanbo, Wang, W., 2019. Inter-comparison of satellite-retrieved and global land data assimilation system-simulated soil moisture datasets for global drought analysis. Remote Sens. Environ. 220, 1–18. https://doi.org/10.1016/j.rse.2018.10.026.

Llamas, R.M., Guevara, M., Rorabaugh, D., Taufer, M., Vargas, R., 2020. Spatial gap-filling of ESA CCI satellite-derived soil moisture based on geostatistical techniques and multiple regression. Remote Sens. 12, 665. https://doi.org/10.3390/rs12040665.

Long, D., Bai, L., Yan, L., Zhang, C., Yang, W., Lei, H., Quan, J., Meng, X., Shi, C., 2019. Generation of spatially complete and daily continuous surface soil moisture of high spatial resolution. Remote Sens. Environ. 233, 111364. https://doi.org/10.1016/j.rse.2019.111364.

Martens, B., Miralles, D.G., Lievens, H., van der Schalie, R., de Jeu, R.A.M., Fernández-Prieto, D., Beck, H.E., Dorigo, W.A., Verhoest, N.E.C., 2017. GLEAM v3: satellite-based land evaporation and root-zone soil moisture. Geosci. Model Dev. 10, 1903–1925. https://doi.org/10.5194/gmd-10-1903-2017.

Martínez-Fernández, J., González-Zamora, A., Sánchez, N., Gumuzzio, A., Herrero-Jiménez, C.M., 2016. Satellite soil moisture for agricultural drought monitoring: assessment of the SMOS derived soil water deficit index. Remote Sens. Environ. 177, 277–286. https://doi.org/10.1016/j.rse.2016.02.064.

Martínez-Fernández, J., Almendra-Martín, L., de Luis, M., González-Zamora, A., Herrero-Jiménez, C., 2019. Tracking tree growth through satellite soil moisture monitoring: a

case study of Pinus halepensis in Spain. Remote Sens. Environ. 235, 111422 https://doi.org/10.1016/j.rse.2019.111422.
Mascaro, G., Vivoni, E.R., Deidda, R., 2011. Soil moisture downscaling across climate regions and its emergent properties. J. Geophys. Res. 116 https://doi.org/10.1029/2011JD016231.
McNallya, A., Shuklab, S., Arsenaultc, K.R., Wangc, S., Peters-Lidardd, C.D., Verdin, J.P., 2016. Evaluating ESA CCI soil moisture in East Africa. Int. J. Appl. Earth Obs. Geoinf. 48, 96–109. https://doi.org/10.1016/j.jag.2016.01.001.
Miralles, D.G., Holmes, T.R.H., De Jeu, R.A.M., Gash, J.H., Meesters, A.G.C.A., Dolman, A.J., 2011. Global land-surface evaporation estimated from satellite-based observations. Hydrol. Earth Syst. Sci. 15, 453–469. https://doi.org/10.5194/hess-15-453-2011.
Mountrakis, G., Im, J., Ogole, C., 2011. Support vector machines in remote sensing: a review. ISPRS J. Photogramm. Remote Sens. 66, 247–259. https://doi.org/10.1016/j.isprsjprs.2010.11.001.
Mutanga, O., Adam, E., Cho, M.A., 2012. High density biomass estimation for wetland vegetation using worldview-2 imagery and random forest regression algorithm. Int. J. Appl. Earth Obs. Geoinf. 18, 399–406. https://doi.org/10.1016/j.jag.2012.03.012.
Nicolai-Shaw, N., Zscheischler, J., Hirschi, M., Gudmundsson, L., Seneviratne, S.I., 2017. A drought event composite analysis using satellite remote-sensing based soil moisture. Remote Sens. Environ. 203, 216–225. https://doi.org/10.1016/j.rse.2017.06.014.
Pal, M., 2005. Random forest classifier for remote sensing classification. Int. J. Remote Sens. 26, 217–222. https://doi.org/10.1080/01431160412331269698.
Pal, M., 2006. Support vector machine-based feature selection for land cover classification: a case study with DAIS hyperspectral data. Int. J. Remote Sens. 27, 2877–2894. https://doi.org/10.1080/01431160500242515.
Peng, J., Loew, A., Merlin, O., Verhoest, N.E.C., 2017. A review of spatial downscaling of satellite remotely sensed soil moisture. Rev. Geophys. 55, 341–366. https://doi.org/10.1002/2016RG000543.
Pérez-Planells, L., Delegido, J., Rivera-Caicedo, J.P., Verrelst, J., 2015. Análisis de métodos de validación cruzada para la obtención robusta de parámetros biofísicos. Rev. Teledetección 44, 55–65. https://doi.org/10.4995/raet.2015.4153.
Piles, M., Sanchez, N., Vall-llossera, M., Camps, A., Martinez-Fernandez, J., Martinez, J., Gonzalez-Gambau, V., 2014. A downscaling approach for SMOS land observations: evaluation of high-resolution soil moisture maps over the Iberian Peninsula. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 7, 3845–3857. https://doi.org/10.1109/JSTARS.2014.2325398.
Piles, M., Petropoulos, G.P., Sánchez, N., González-Zamora, Á., Ireland, G., 2016. Towards improved spatio-temporal resolution soil moisture retrievals from the synergy of SMOS and MSG SEVIRI spaceborne observations. Remote Sens. Environ. 180, 403–417. https://doi.org/10.1016/j.rse.2016.02.048.
Qiu, J., Gao, Q., Wang, S., Su, Z., 2016. Comparison of temporal trends from multiple soil moisture data sets and precipitation: the implication of irrigation on regional soil moisture trend. Int. J. Appl. Earth Obs. Geoinf. 48, 17–27. https://doi.org/10.1016/j.jag.2015.11.012.
Qu, Y., Zhu, Z., Chai, L., Liu, S., Montzka, C., Liu, J., Yang, X., Lu, Z., Jin, R., Li, X., Guo, Z., Zheng, J., 2019. Rebuilding a microwave soil moisture product using random Forest adopting AMSR-E/AMSR2 brightness temperature and SMAP over the Qinghai–Tibet plateau, China. Remote Sens. 11, 683. https://doi.org/10.3390/rs11060683.
Rigling, B.D., 2012. Application of temporal gap filling to natural power law spectrums. IEEE Geosci. Remote Sens. Lett. 9, 624–628. https://doi.org/10.1109/LGRS.2011.2177062.
Sakai, T., Iizumi, T., Okada, M., Nishimori, M., Grünwald, T., Prueger, J., Cescatti, A., Korres, W., Schmidt, M., Carrara, A., Loubet, B., Ceschia, E., 2016. Varying applicability of four different satellite-derived soil moisture products to global gridded crop model evaluation. Int. J. Appl. Earth Obs. Geoinf. 48, 51–60. https://doi.org/10.1016/j.jag.2015.09.01.
Sandholt, I., Rasmussen, K., Andersen, J., 2002. A simple interpretation of the surface temperature/vegetation index space for assessment of surface moisture status. Remote Sens. Environ. 79, 213–224. https://doi.org/10.1016/S0034-4257(01)00274-7.
Seneviratne, S.I., Corti, T., Davin, E.L., Hirschi, M., Jaeger, E.B., Lehner, I., Orlowsky, B., Teuling, A.J., 2010. Investigating soil moisture–climate interactions in a changing climate: a review. Earth-Science Rev. 99, 125–161. https://doi.org/10.1016/j.earscirev.2010.02.004.
Smola, A.J., Schölkopf, B., 2004. A tutorial on support vector regression. Stat. Comput. 14, 199–222. https://doi.org/10.1023/B:STCO.0000035301.49549.88.
Soil Moisture Climate Change Initiative (CCI), 2020. https://www.esa-soilmoisture-cci.org (accessed 22 June 2020).
Srivastava, P.K., Han, D., Ramirez, M.R., Islam, T., 2013. Machine learning techniques for downscaling SMOS satellite soil moisture using MODIS land surface temperature for hydrological application. Water Resour. Manag. 27, 3127–3144. https://doi.org/10.1007/s11269-013-0337-9.
Tallec, G., Ansart, P., Guérin, A., Delaigue, O., Blanchouin, A., 2015. Observatoire Oracle. Irstea. https://dx.doi.org/10.17180/obs.oracle.
Vapnik, V.N., 1995. The Nature of Statistical Learning Theory. Springer New York, New York, NY. https://doi.org/10.1007/978-1-4757-3264-1.
Wang, G., Kim, Y., Wang, D., 2007. Quantifying the strength of soil moisture-precipitation coupling and its sensitivity to changes in surface water budget. J. Hydrometeorol. 8, 551–570. https://doi.org/10.1175/JHM573.1.
Wang, G., Garcia, D., Liu, Y., de Jeu, R., Johannes Dolman, A., 2012. A three-dimensional gap filling method for large geophysical datasets: application to global satellite soil moisture observations. Environ. Model. Softw. 30, 139–142. https://doi.org/10.1016/j.envsoft.2011.10.015.
Wang, S., Mo, X., Liu, S., Lin, Z., Hu, S., 2016. Validation and trend analysis of ECV soil moisture data on cropland in North China plain during 1981–2010. Int. J. Appl. Earth Obs. Geoinf. 48, 110–121. https://doi.org/10.1016/j.jag.2015.10.010.
Wang, T., Franz, T.E., Li, R., You, J., Shulski, M.D., Ray, C., 2017. Evaluating climate and soil effects on regional soil moisture spatial variability using EOFs. Water Resour. Res. 53, 4022–4035. https://doi.org/10.1002/2017WR020642.
Xiao, Z., Jiang, L., Zhu, Z., Wang, J., Du, J., 2016. Spatially and temporally complete satellite soil moisture data based on a data assimilation method. Remote Sens. 8 https://doi.org/10.3390/rs8010049.
Xie, X., Liu, W.T., Tang, B., 2008. Spacebased estimation of moisture transport in marine atmosphere using support vector regression. Remote Sens. Environ. 112, 1846–1855. https://doi.org/10.1016/j.rse.2007.09.003.
Xing, C., Chen, N., Zhang, X., Gong, J., 2017. A machine learning based reconstruction method for satellite remote sensing of soil moisture images with *in situ* observations. Remote Sens. 9 https://doi.org/10.3390/rs9050484.
Yang, F., White, M.A., Michaelis, A.R., Ichii, K., Hashimoto, H., Votava, P., Zhu, A.X., Nemani, R.R., 2006. Prediction of continental-scale evapotranspiration by combining MODIS and AmeriFlux data through support vector machine. IEEE Trans. Geosci. Remote Sens. 44, 3452–3461. https://doi.org/10.1109/TGRS.2006.876297.
Zacharias, S., Bogena, H., Samaniego, L., Mauder, M., Fuß, R., Pütz, T., Frenzel, M., Schwank, M., Baessler, C., Butterbach-Bahl, K., Bens, O., Borg, E., Brauer, A., Dietrich, P., Hajnsek, I., Helle, G., Kiese, R., Kunstmann, H., Klotz, S., Munch, J.C., Papen, H., Priesack, E., Schmid, H.P., Steinbrecher, R., Rosenbaum, U., Teutsch, G., Vereecken, H., 2011. A network of terrestrial environmental observatories in Germany. Vadose Zo. J. 10, 955–973. https://doi.org/10.2136/vzj2010.0139.
Zhang, X., Chen, N., 2016. Reconstruction of GF-1 soil moisture observation based on satellite and *In Situ* sensor collaboration under full cloud contamination. IEEE Trans. Geosci. Remote Sens. 54, 5185–5202. https://doi.org/10.1109/TGRS.2016.2558109.
Zhao, W., Sánchez, N., Lu, H., Li, A., 2018. A spatial downscaling approach for the SMAP passive surface soil moisture product using random forest regression. J. Hydrol. 563, 1009–1024. https://doi.org/10.1016/j.jhydrol.2018.06.081.