scientific data

Check for updates

OPEN

DATA DESCRIPTOR

High-resolution European daily soil moisture derived with machine learning (2003–2020)

Sungmin O[1] ✉, Rene Orth[2], Ulrich Weber[2] & Seon Ki Park[1,3,4] ✉

Machine learning (ML) has emerged as a novel tool for generating large-scale land surface data in recent years. ML can learn the relationship between input and target, e.g. meteorological variables and *in-situ* soil moisture, and then estimate soil moisture across space and time, independently of prior physics-based knowledge. Here we develop a high-resolution (0.1°) daily soil moisture dataset in Europe (SoMo.ml-EU) using Long Short-Term Memory trained with *in-situ* measurements. The resulting dataset covers three vertical layers and the period 2003–2020. Compared to its previous version with a lower spatial resolution (0.25°), it shows a closer agreement with independent *in-situ* data in terms of temporal variation, demonstrating the enhanced usefulness of *in-situ* observations when processed jointly with high-resolution meteorological data. Regional comparison with other gridded datasets also demonstrates the ability of SoMo.ml-EU in describing the variability of soil moisture, including drought conditions. As a result, our new dataset will benefit regional studies requiring high-resolution observation-based soil moisture, such as hydrological and agricultural analyses.

Background & Summary

Because soil moisture is a key variable in water, energy, and biogeochemical cycles[1], the availability of large-scale, high-resolution soil moisture datasets can facilitate diverse weather and climate applications, including monitoring and forecasting of hydrological and ecological extreme events[2–7]. Soil moisture exhibits complex spatial and temporal dynamics, which undermines the usefulness of the direct use of sparse *in-situ* measurements for estimations or even predictions across larger areas[8,9]. Consequently, different approaches have been explored as a way of obtaining improved spatial data coverage. Physics-based land surface models have played the main role in providing continental-to-global scale soil moisture datasets[10–12]. Despite the fact that soil moisture values are not comparable and transferable among models[13], temporally and spatially continuous data can be obtained across multiple soil layers. Satellite observations are another important resource, providing soil moisture estimates across large areas[14–17], although their data are often missing due to the local retrieval conditions (e.g. presence of clouds or dense vegetation) or satellite revisit times. Moving beyond model and satellite-based datasets, novel machine learning (ML) approaches have been increasingly employed in recent years to generate large-scale soil moisture datasets, for instance, by filling the temporal and spatial gaps in satellite observations or by integrating multiple data sources[18–20].

In this context, our previous study[21] (hereafter O21) used ML to generate a global soil moisture dataset – SoMo.ml – by upscaling *in-situ* measurements. We trained ML with meteorological forcing and soil moisture, such that the ML could learn emerging input-output relationships and thereby establish its own knowledge about processes. In this way, ML provides an attractive avenue to provide data that are independent from traditional modeling or satellite-based datasets. This is particularly useful for the regions where established approaches can not provide reliable estimates, e.g. high latitudes or densely-vegetated areas, because with another independent soil moisture data source, we can mitigate data uncertainty by comparing or using an ensemble of diverse datasets generated by different approaches[22,23]. Further, ML-driven datasets can foster scientific discovery in the Earth and climate domain dominated by physics-based models, particularly for processes that are not

[1]Department of Climate & Energy System Engineering, Ewha Womans University, Seoul, Korea. [2]Department of Biogeochemical Integration, Max Planck Institute for Biogeochemistry, Jena, Germany. [3]Center for Climate/Environment Change Prediction Research, Ewha Womans University, Seoul, Korea. [4]Severe Storm Research Center, Ewha Womans University, Seoul, Korea. ✉e-mail: sungmin.o@ewha.ac.kr; spark@ewha.ac.kr

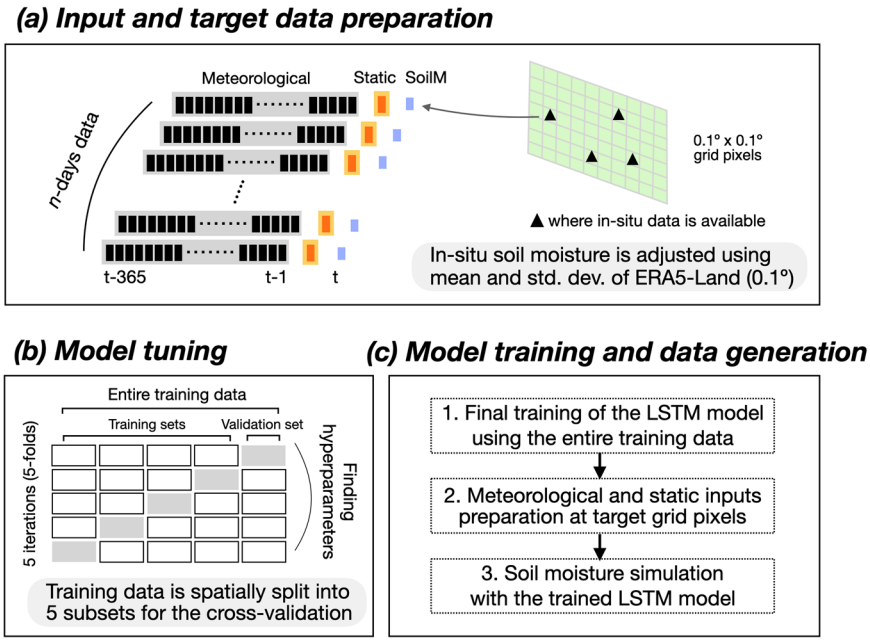

Fig. 1 Schematic overview of the workflow. (**a**) Meteorological variables and static features are collected at the grid pixels where *in-situ* soil moisture measurements exist. The *in-situ* soil moisture data are adjusted to match the long-term mean and standard deviation of respective ERA5 gridded soil moisture data; note that daily variations are directly from the *in-situ* measurements. (**b**) Cross-validation is performed using the training data spatially partitioned into five subsets, and optimal hyperparameters are determined based on the performance mean of the five validation sets. (**c**) The final LSTM model with the optimal hyperparameters is trained with the entire training data. The trained model is then used to estimate soil moisture at every target grid pixel within the domain.

sufficiently well represented in the models[24,25]. For instance, SoMo.ml has contributed to observation-based analysis of drought-related ecosystem damages or heat events[26–28].

Despite the availability of various data sources, acquiring high-resolution soil moisture data remains a challenge[29,30]. Various downscaling approaches have been developed to improve the spatial resolution of existing data[31], however, the publicly available large-scale datasets are mostly available at coarse resolutions (25–50 km). To better reflect the land surface heterogeneity associated with soil and vegetation characteristics and, thereby, improve understanding of soil moisture heterogeneity and related hydrological and meteorological processes, there is a need for a more detailed picture of soil moisture dynamics across a broad range of relevant communities[30,32–34].

Given the wide range of applications for soil moisture data, and the increasing need for high-resolution data, we aim to provide an updated ML-driven soil moisture dataset for Europe and document the added value of increased spatial resolution. To do this, we largely follow the methodology of O21 (Fig. 1); Long Short-Term Memory (LSTM) networks ingest meteorological forcings and static features as inputs and return estimated soil moisture as outputs (targets). LSTM is a special kind of recurrent neural network[35], designed specifically to overcome the long-range temporal dependency problem of the traditional recurrent neural networks. This makes LSTM networks well suited to model soil moisture which, as a state variable and given the soil water holding capacity, integrates the meteorological forcing and thereby is driven by both concurrent and preceding meteorology[21,36,37]. We scale point-level *in-situ* data to grid-scale target data by adjusting the mean and standard deviation of the *in-situ* data to those of ERA5-Land reanalysis 0.1° gridded soil moisture[12] within the overlapping time period. In other words, the long-term mean and variation of the target data largely follow ERA5-Land, while daily variation is retrieved from *in-situ* measurements. As a result, our new dataset SoMo.ml-EU provides multi-layer daily soil moisture (0–10, 10–30, and 30–50 cm depths) for Europe at a 0.1° resolution. The resulting data show comparable performance to other widely-used gridded datasets. The most distinct feature of SoMo.ml-EU is that, by design, it closely follows daily variations in *in-situ* data, even more so than the downscaled previous version. Consequently, SoMo.ml-EU will be useful for various studies and applications that require high-resolution, observation-based soil moisture information.

Methods

Training data preparation. *Target In-situ soil moisture.* We obtain daily *in-situ* soil moisture data from the International Soil Moisture Network (ISMN). ISMN is an international data hosting facility that collects, harmonises, and shares *in-situ* soil moisture measurements available around the globe[38,39]. We consider *in-situ* data not only from the European domain but also from the US. In this way, we are able to obtain a greater amount of training data, which also represents more diverse climate conditions, as shown in Fig. 2. See also Fig. S1. The

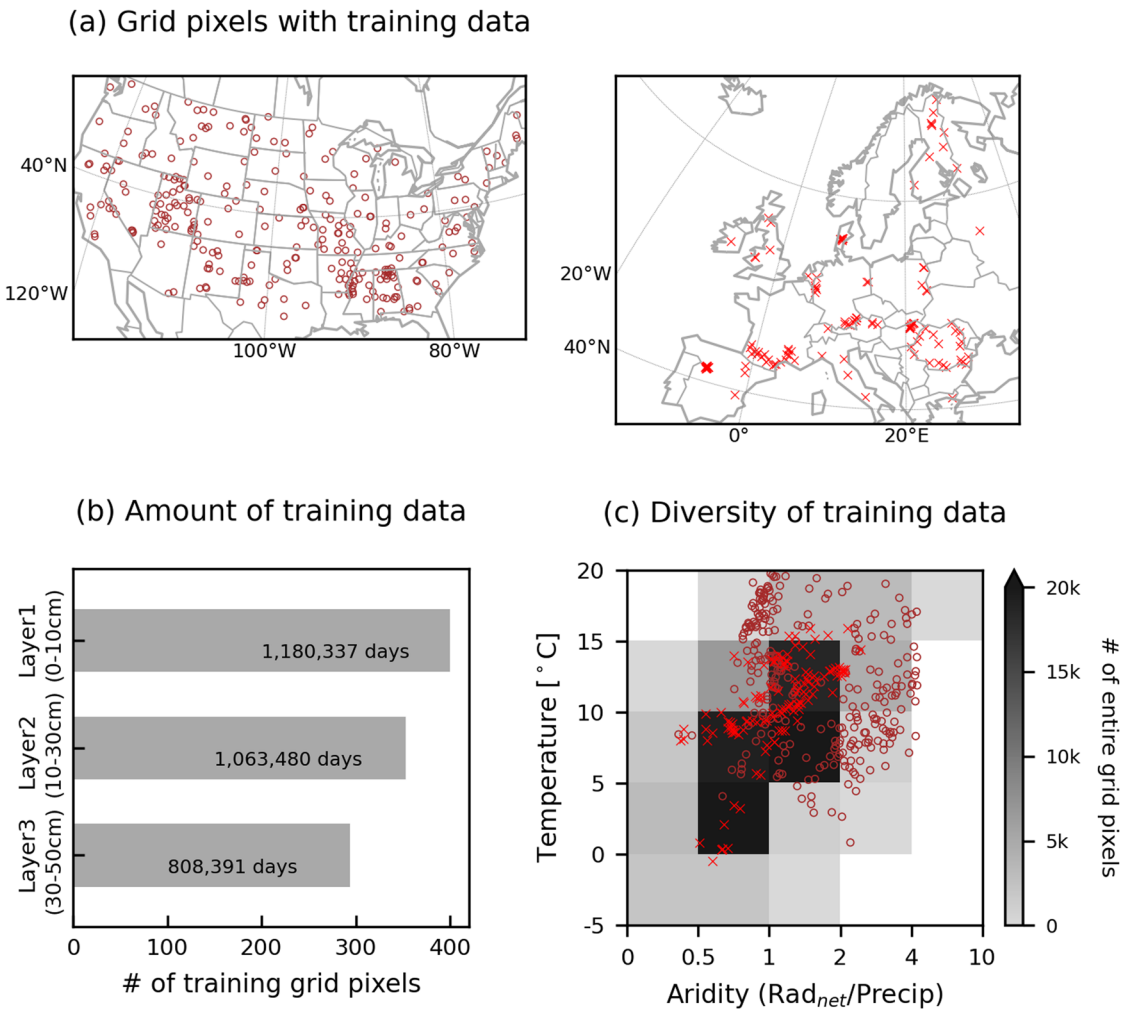

Fig. 2 (**a**) Spatial distribution of training data; brown circles and red markers show the grid pixels where *in-situ* soil moisture data are obtained in US and Europe, respectively. (**b**) The number of grid pixels and the total amount of training data for each layer. (**c**) Hydroclimatic diversity of origins of training data for the first layer; markers show the training data distribution across different climatic regimes defined by the long-term mean aridity and temperature of each grid pixel. Grey-scale colours represent the number of grid pixels for each regime across the European domain that SoMo.ml-EU covers. The same information for the second and third layers can be found in Fig. S1.

LSTM-based model is trained using a combination of training data from all available locations; it is possible because (i) the ML model doesn't require the geographical information of training data, but it ingests the entire training data jointly, and (ii) ML can transfer knowledge (the input-output relationships) across space[40], which is often the issue to physically-based models (i.e. parameter regionalisation[41]). In fact, more data helps the model to accurately describe input-output relationships and more diverse data enables the model to represent the variation of these relationships across conditions. The resulting accurate relationships, and their variations across conditions, form the backbone of the LSTM-based estimation of soil moisture in different time periods and regions. While the capacity of ML models to extrapolate input-output relationships beyond known conditions is limited, they can transfer knowledge between space and time at the same time[40]; e.g. the models learn about the input-output relationships in less-sampled semi-arid sites with many training data collected during dry periods at humid sites. Table 1 shows a full list of the ISMN networks contributing to the training data.

The collected *in-situ* data are then scaled using the long-term mean and standard deviation of ERA-Land gridded soil moisture. This scaling is intended to eliminate systematic biases in soil moisture means and variabilities among different types of sensors and calibration techniques while maintaining the daily variations observed directly by *in-situ* measurements, consequently yielding upscaled soil moisture information at the target resolution of 0.1°. It is worth noting that temporal variations in point-level data have a wider footprint than absolute soil moisture values[42]. If there are more than one scaled measurement time series for a target grid pixel, we take an average of them. See O21 for further details about the data preparation.

Network	Country	Data aquisition
BIEBRZA-S-1[61]	Poland	Layer1/2/3
CALABRIA	Italy	Layer-/2/-/
CAMPANIA[62]	Italy	Layer-/2/-/
COSMOS[63]	USA	Layer1/2/-/
FMI[64]	Finland	Layer1/2/3
FR-Aqui[65]	France	Layer1/2/3
GROW[66]	UK	Layer1/-/-/
GTK	Finland	Layer1/2/3/
HOAL	Austria	Layer1/2/3/
HOBE[67]	Denmark	Layer1/2/-/
HYDROL-NET-PERUGIA[68]	Italy	Layer1/2/3/
IMA-CAN1[69]	Italy	Layer1/-/-/
IPE	Spain	Layer1/2/-/
METEROBS	Italy	Layer1/2/3/
MOL-RAO[70]	Germany	Layer1/2/3/
REMEDHUS[71]	Spain	Layer1/-/-/
RSMN	Romania	Layer1/-/-/
Ru-CFR	Russia	Layer1/2/3/
SCAN[72]	USA	Layer1/2/3/
SMOSMANIA[73]	France	Layer1/2/-/
SWEX-POLAND[74]	Poland	Layer1/2/-/
TERENO[75]	Germany	Layer1/2/3/
UDC-SMOS[76]	Germany	Layer1/2/3/
UMBRIA[62]	Italy	Layer-/2/3/
UMSUOL	Italy	Layer1/2/3/
USCRN[77]	USA	Layer1/2/3/
VAS	Spain	Layer1/-/-/
WEGENERNET[78]	Austria	Layer-/2/-/
WSMN[79]	UK	Layer1/-/-/

Table 1. List of the ISMN networks considered to collect *in-situ* measurements.

Meteorological forcing variables and static features. We use temperature, precipitation, net radiation, and skin temperature (land surface temperature) as meteorological forcing variables for the first layer. Each layer is computed with a separate LSTM. The model first predicts surface soil moisture using the four meteorological variables as inputs. Then, the models for the second and third layers additionally use the soil moisture predicted from the upper layer(s). These variables were found to be the most influential variables according to our previous analysis in O21, which quantified the contribution of the respective variables to the LSTM model performance. Other relevant variables, such as specific humidity or wind, only had a marginal effect on the performance of the LSTM; moreover, it remains challenging to obtain reliable high-resolution gap-free and observation-based data for these variables over large areas.

The previous version of SoMo.ml used only ERA5 reanalysis data, whereas the meteorological data for this study are pulled from more diverse data sources (see Table 2) in order to reduce the dependency of the resulting data product on ERA5. Forcing data with a sub-daily resolution are aggregated to a daily scale, and all data are available at 0.1° resolution such that no spatial aggregation is applied. We examine the feature importance with the input variables employed in this study (Fig. S2). The results are similar to those found in O21, despite using different data sources. That is, precipitation and skin temperature are the most important meteorological inputs in surface soil moisture simulations, while the upper layer's soil moisture is most important in deeper layer simulations.

We also use static features such as climatological values (long-term mean precipitation and aridity), topography (means and standard deviations of sub-grid scale elevation values), vegetation (forest and short vegetation fractions), and soil properties (sand and clay fractions for surface and deep layers). These static inputs have proven to be particularly important in surface soil moisture simulations[21] (Fig. S2). We prepare all the static data at the spatial resolution of the target data (i.e. $0.1° \times 0.1°$). For instance, we compute the mean and standard deviation of sub-grid scale (1 arc-minute) elevations[43] within the target resolution for topography information. For the vegetation and soil information[44], we compute the mean of sub-grid scale values. A comprehensive description of the meteorological forcing and static datasets is given in Table 2.

Model training and application. We employ a modified version of LSTM architecture, Entity-Aware LSTM[45], that can explicitly differentiate between dynamic and static inputs, i.e. meteorological forcing and static features in our case. The inputs are normalised by subtracting the mean and dividing by the standard deviation to stabilise and accelerate training[46]. We determine the optimal set of hyperparameter values, which govern the

	Variable	Source	Description
Dynamic	Air temperature	MSWX-Past[80]	Bias-corrected and downscaled Climatic Research Unit Time Series data
	Precipitation	MSWEP[81]	Gauge, satellite, and reanalysis merged data
	Net surface radiation	ERA5-Land[12]	ECMWF land reanalysis
	Skin temperature	ERA5-Land[12]	ECMWF land reanalysis
	Soil moisture from upper layer(s)	SoMo.ml-EU	ML-based soil moisture produced in this study
Static	Mean precipitation	MSWEP[81]	Long-term mean precipitation
	Aridity	MSWEP[81], ERA5-Land[12]	Ratio of net radiation (unit converted to *mm*) to precipitation
	Topography	ETOPO1[43]	Mean and standard deviation of sub-grid scale elevations
	Land cover	Harmonized World Soil Database v1.2[82]	Forest and short vegetation fraction computed based on six geographic datasets
	Soil type	Harmonized World Soil Database v1.2 (regridded)[44]	Clay and sand fraction based on regional and national updates of soil information

Table 2. Description of meteorological forcing variables and static features.

model structures and training processes, through k-fold cross-validation with $k = 5$ (Fig. 1). The final model has 128 hidden units in one LSTM layer with one dense layer, and the dropout rate is set to 0.4. Another important hyperparameter is the length of the input sequence (i.e. lookback), corresponding to the number of prior time steps of meteorological data, e.g. from t-365 to t-1, used to predict soil moisture at a time step t. The lookback is typically set to 365 days in daily hydrologic simulations with an assumption that the LSTM model can learn the dynamics of an entire preceding annual cycle[47]. In order to better specify this hyperparameter, and to study soil moisture memory characteristics in LSTM, we conduct an additional experiment to assess the effect of input sequence length on model performance as part of the model validation process (see the following section).

Finally, we train the LSTM model using the entire training dataset. The trained model with all established relationships between the input variables and soil moisture across concurrent and previous times is then applied across the entire European grid pixels. We repeat the simulations five times and take the averages of the simulations to compute the final soil moisture given the random initialisation of LSTM weights (learnable parameters).

Data Records

SoMo.ml-EU provides volumetric soil moisture over the domain of longitudes 12.0 °W to 45.0 °E and latitudes 36.0 °N to 71.5 °N. The data covers the period 2003–2020, and the spatiotemporal resolution is 0.1° and one day. The data files are freely available at figshare[48]. An example file name is 'SoMo.ml-EU_*LAYER*_*YYYY*.nc', with *LAYER* and *YYYY* standing for soil moisture layer depth and year, respectively.

Technical Validation

We report the suitability of LSTM-based model for soil moisture simulations, focusing on two aspects. First, we assess the impact of preceding meteorological information on modelling performance. As mentioned, it is a key characteristic of LSTM networks to consider the relationship of the target variable with both concurrent and preceding input variable estimates. Second, we evaluate the ability of LSTM to spatially extrapolate learned input-output relationships. This knowledge transfer between locations is critical to generating enhanced quality data in domains with limited training data.

Furthermore, we report the performance of the final data product of SoMo.ml-EU through intercomparison with multiple gridded datasets. We employ (1) SoMo.ml v1, the previous version of SoMo.ml with a lower spatial resolution (0.25°), (2) ERA5-Land, a replay of the land component of the ERA5 reanalysis[12], (3) GLEAM satellite-based data; we choose v3.5.b, which is based on satellite data only and no reanalysis is used[49], and (4) CLM-DA, high-resolution soil moisture generated from the Community Land Model with data assimilation of ESA CCI satellite observations[11]. Finally, we compare the datasets using independent *in-situ* soil moisture data obtained from the COSMOS-Europe network[50]. It is a newly introduced soil moisture network in Europe, which offers the opportunity for an independent and comparative evaluation as these measurements are not used in the derivation of any of the gridded products we compare here. Additionally, we compare the dataset at locations of around 300 grid pixels, selected from every 20×20 pixels segment over the entire domain, to overcome the limitation of the low climate diversity represented by the COSMOS-Europe data, using the average of the considered datasets as a reference.

Role of preceding meteorological information. First, we train the LSTM with input data covering the preceding 365 days and define its prediction performance as 'reference performance'. Next, we randomly permuted a part of the sequences except for the last n time steps, i.e. from t-365 to t-$(n+1)$ days, where t is the prediction time step. Consequently, only the later subsequent part of the time series, i.e. from t-n to t-1 days, keeps

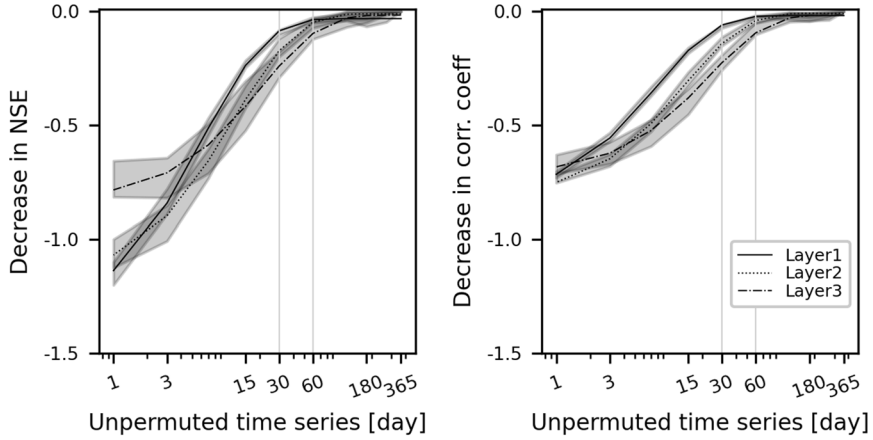

Fig. 3 Importance of preceding meteorological information for the performance of the resulting soil moisture simulation. This is expressed as performance reduction when perturbing part of the preceding meteorological information, compared with the reference model performance using unchanged meteorological information. For example, results for 30 days are obtained with meteorological input data left unchanged for the last 30 days while being permuted for the first 365-30 = 335 days. Nash–Sutcliffe efficiency and correlation coefficient are considered for performance evaluation. We repeat the simulations with different hyperparameters related to model architecture, e.g. the number of layers or hidden units. The shaded area shows the 0.2 to 0.8 quantiles of the metrics across model architectures. Note that x-axes use a logarithmic scale.

its original information. Figure 3 shows the model performance for the surface soil moisture simulation using the partially permuted input sequences. The model performance significantly decreases when the unperturbed portion of the input time series becomes shorter than 30 to 60 days. On the other hand, the meteorological data prior to the preceding two months show a small contribution to the LSTM performance. This matches well with previous studies quantifying soil moisture memory time scales around several weeks to months, albeit with seasonal and regional variations[51,52]. We repeat the simulations for the deeper layers and find a weaker model performance for the same degree of input data perturbation such that a similar model performance requires more preceding input information compared to the first layer simulations. This implies longer soil moisture memory in the deeper layers, which is consistent with the findings from analyses that used physics-based model data[53].

For the final model simulation, we keep the input sequence length (lookback) to 365 days to exploit the full potential of the meteorological information, i.e. giving the model the flexibility to account for potentially longer memory time scales occurring in specific regions and times. At the same time, these results reported in Fig. 3 have important implications. First, LSTM effectively extracts useful information from prior time steps and this explicitly contributes to more accurate soil moisture simulations. Therefore, a sufficient lookback period should always be considered. Second, we show the delayed effect of meteorological input on soil moisture and, therefore, the usefulness of LSTM to quantify soil moisture memory. Future research can focus on determining seasonal and regional variations of the soil moisture memory, potentially offering new insights into the conventional methods, e.g. using autocorrelation-based metrics[52,54].

Model performance across continents. We train the model with training data from the US domain only and validate the model performance over the European domain. In this way, we can evaluate the model with an emphasis on its capability to transfer knowledge across continents, and significantly minimise the risk of overfitting due to spatial autocorrelation between closely located training and validation data. The simulated soil moisture here is referred to as SoMo.ml* (Fig. 4), because they are different from the final SoMo.ml-EU data generated from the model trained with the entire training data from the US and Europe both. We also report the model validation results from the five-fold cross-validation in Fig. S3 in the Supplementary material. Note that we split the entire training data *spatially* to prepare the training and validation subsets. It is because the spatial extrapolation capacity of ML is more critical for our data generation, i.e. estimating soil moisture over grid pixels with no training data. Nonetheless, we split the training data *temporally* and again examine the performance of ML with the selected hyperparameters. The results confirm that the temporal predictability of ML is comparable to the spatial extrapolation performance of ML (Fig. S4).

Overall, the model shows satisfactory performance (Fig. 4). The results clearly show how efficiently ML can transfer knowledge between locations, although relatively poorer performance is observed in Layer 1 (0–10 cm depth), which is mainly due to outliers, i.e. target soil moisture from the five grid pixels with a very humid climate where the average soil moisture is above 0.5 m^3/m^3. This illustrates the main limitation of ML approaches. They can hardly extrapolate beyond the conditions covered by the training data. The US data are mostly sampled from arid and warm regions, so training data for humid and cold regions are lacking (see Fig. 2(c)). The model shows a better performance for Layer 2 and Layer 3 (10–30 cm and 30–50 cm depths, respectively), which could be related to the lower temporal variability or to the relatively simple input-output relationships (deeper layer soil moisture is mainly determined by the upper layer soil moisture than dynamic meteorological information[21])

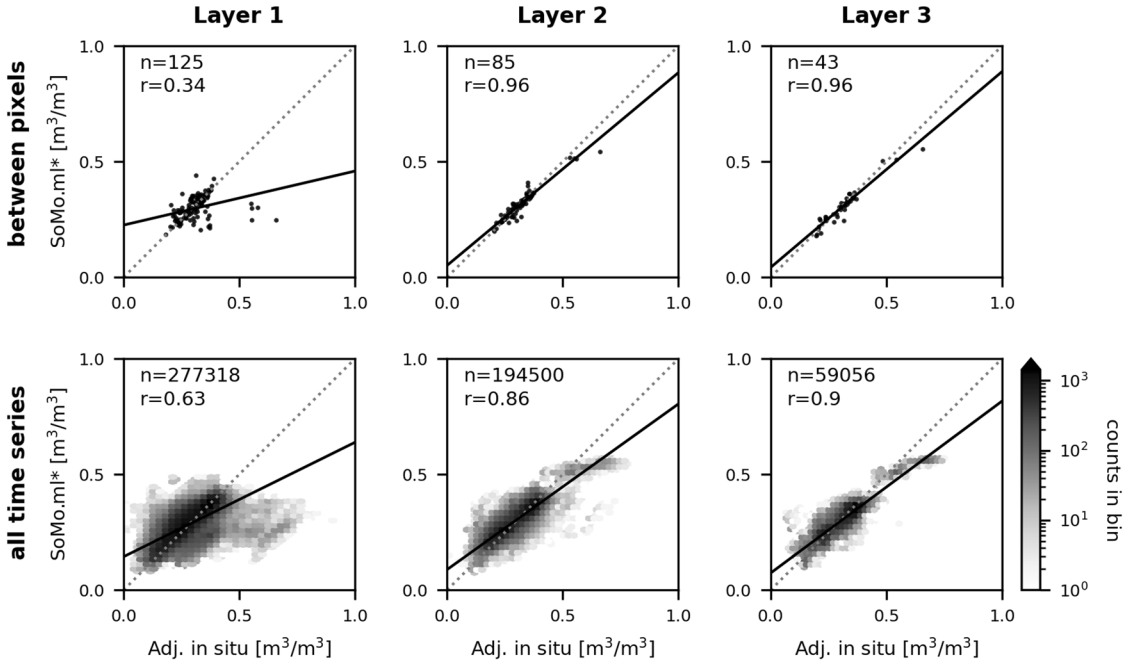

Fig. 4 Transfer learning capacity of the LSTM network. The model performance is validated in Europe after training with data from the US only. We compare pixel-averaged soil moisture (top) and daily soil moisture from all grid pixels (bottom) in Europe between target soil moisture and SoMo.ml* at each layer. See Figs. S3 and S4 for model performance evaluated through the k-fold cross-validation.

in the deeper layers that can be relatively easily reproducible by the model. Note that the model performance for wet regions is better ($r = 0.8$ for correlation between pixels in Layer 1) according to the five-fold cross-validation in which the training data from Europe are included (Fig. S3). The final model is trained with the entire training data and, therefore, the performance of the actual SoMo.ml-EU over humid regions (similar to the training conditions at high latitudes of Europe) is expected to be reasonable.

Comparison against independent *in-situ* measurements. In Fig. 5a, we compare SoMo.ml-EU, including multiple gridded datasets, against *in-situ* data obtained from the COSMOS-Europe network[50]. We select validation grid pixels that are not already included in our training data, although the COSMOS-Europe data only cover a relatively narrow range of climatic regimes. Note that COSMOS-EU does not include consistent data from all depths at all sites (see Fig. S1). The datasets with different native spatial resolutions are regridded to 0.1° using bilinear interpolation. We find that SoMo.ml-EU agrees closer with the *in-situ* data than the other gridded datasets including the previous SoMo.ml version, confirming the benefit of constructing high-resolution data rather than simply using interpolated data.

In terms of unbiased root mean square difference (uRMSD), SoMo.ml-EU shows smaller deviations from the *in-situ* data, while the median range of uRMSD across the datasets stays narrow between 0.04 to 0.05. SoMo.ml-EU and the other datasets all exhibit lower performance in Layer 2 than Layer 1, because the *in-situ* reference data are collected from more diverse climate conditions (see Fig. S1); reliable soil moisture derivations for extreme regimes are challenging to all approaches. In Fig. 5b, the data are furthermore compared against the average of all the gridded datasets over the randomly selected grid pixels (Fig. S5) from which we do not have independent *in-situ* reference data. SoMo.ml-EU gives a reasonable performance within the range of the other datasets. Interestingly, ERA5-Land shows the highest agreement with the averages, which might be attributable to the dependency of the other datasets on ERA reanalyses. For instance, SoMo.ml used ERA5 meteorological data as the input and, for CLM-DA, ERA-Interim data is involved in the process of forcing data generation.

Comparison of soil moisture variability. In this section, we assess the ability of SoMo.ml-EU to capture soil moisture variability, focusing on drought events. Figure 6(a) presents the spatial distribution of the average surface soil moisture from SoMo-EU and the other gridded datasets during summer season (JJA). Overall, the spatial soil moisture patterns are similar across all datasets, with southwest to northeast gradients from dry to wet. Nonetheless, regional differences are observed, such as in mountain areas or high latitudes where observational data are typically lacking or highly uncertain due to, for instance, soil freezing. Therefore, high uncertainty is expected both for data-driven and observation-assimilated models. The spatial correlation (Pearson's r) between SoMo.ml-EU and the other datasets ranges from 0.48 to 0.56 for the entire domain, whereas the correlation increases to 0.63–0.75 when the latitudinal domain higher than 60 °N is excluded. Even though we find good agreement, in terms of the spatial patterns, large deviations in the absolute values exist with the soil moisture from CLM-DA being especially generally underestimated. This dry bias can be due to a discrepancy in the definitions

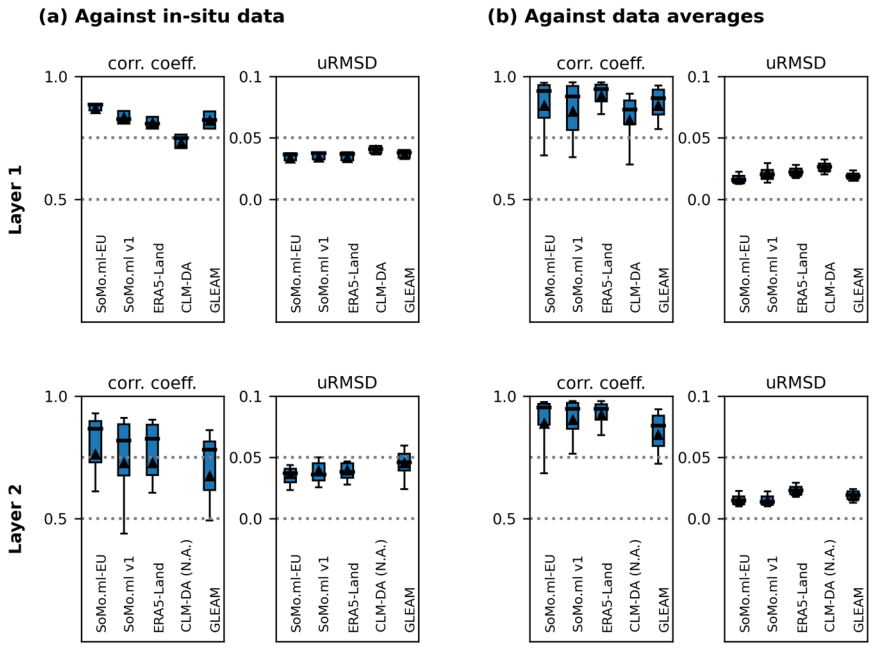

Fig. 5 (**a**) Comparison of soil moisture datasets against reference *in-situ* data over 6 and 22 grid pixels for Layers 1 and 2, respectively, and (**b**) against the average of all datasets at 287 randomly selected grid pixels from across the domain. Gridded soil moisture fields from SoMo.ml-EU, SoMo.ml v1, ERA-Land, CLM-DA, and GLEAM are considered. The climate diversity represented by the reference data used here is shown in Figs. S1 and S5. Correlation coefficient and unbiased root mean square difference are computed for the comparison. Triangles show means and box plot whiskers indicate the 0.2 to 0.8 quantiles of the metrics across grid pixels. See Fig. S7 for results for Layer 3.

of the surface layer[11], e.g. 10 cm for SoMo.ml versus 3 cm for CLM-DA, or the use of observational data, e.g. *in-situ* measurements in SoMo.ml, while satellite data in CLM-DA.

We further compare the normalised soil moisture anomalies as depicted in Fig. 6(b) during 2015 when large parts of Europe were affected by drought[55,56]. The normalised anomalies are calculated by subtracting the long-term seasonal mean and then dividing by the standard deviation (z-score). The datasets show high agreement with similar patterns of positive and negative anomalies. The spatial correlation values range from 0.74 to 0.91. The spatial extent of the affected area described by SoMo.ml-EU, i.e. pronounced negative anomalies over Central and Eastern Europe, matches well with the other datasets, except for the CLM-DA, which shows relatively weaker anomalies.

We also compare seasonal anomalies over the entire overlapping years for different parts of Europe. The results show (Fig. S6) consistent temporal variations between SoMo.ml-EU and the other datasets, particularly for Southern Europe. On the other hand, for Northern Europe, we find relatively large deviations across the datasets, probably because of the high uncertainty in both observational data and physical model representation relating to freeze-thaw processes. More specific examples of temporal variability of SoMo.ml-EU can be found in Fig. 7. We randomly select three different pixels, where training data are available for all three layers, and compare time series of soil moisture between the training (target) and predicted data. The predicted SoMo.ml-EU (dotted lines in the figure) closely follows the temporal pattern of the target soil moisture (solid lines), although data discrepancies are observed, especially when soil moisture gradually decreases. Overall temporal behaviour of SoMo.ml-EU seems reasonable with respect to the temporal variations of precipitation and skin temperature (two most important inputs), even if there are insufficient training data from the same grid pixel (e.g. Pixel 3), which is enabled through the ML's spatial extrapolation.

In general, the soil moisture in SoMo.ml-EU shows similar spatial patterns and absolute values to those of ERA5-Land, which is no surprise given that ERA5-Land data are involved in the training data adjustment (Sect. Training data preparation). The interesting point here is, however, that despite the absence of training data for most of the grid pixels, SoMo.ml-EU shows reasonable spatial patterns and is in general agreement with the other datasets. This demonstrates how efficiently LSTM can spatially extrapolate target variables over large regions, which is enabled by our training strategy of using a combination of training data obtained from diverse climate regimes. This way, these results illustrate the usefulness of using *in-situ* soil moisture data from outside the European target domain to increase the amount and diversity of training data while benefiting from the transfer learning ability of the LSTM networks.

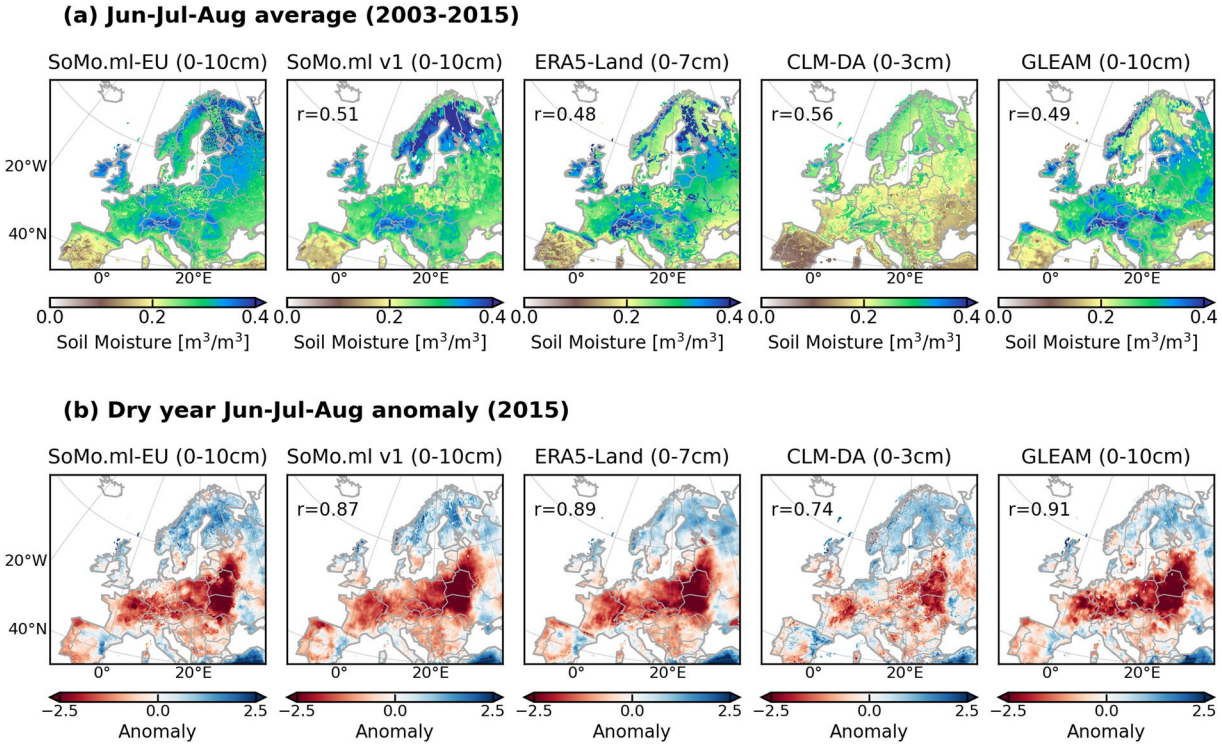

Fig. 6 Comparison of mean and extreme soil moisture from SoMo.ml-Eu with other state-of-the-art datasets; spatial distribution of (**a**) the average summer (JJA) soil moisture between 2003 and 2015 and (**b**) the normalised summer soil moisture anomaly in 2015. Pearson's correlation of the spatial patterns (r) is computed between SoMo.ml-EU and the other data. All datasets are regridded to 0.1 degrees. Results for Layers 2 and 3 are provided in Fig. S8.

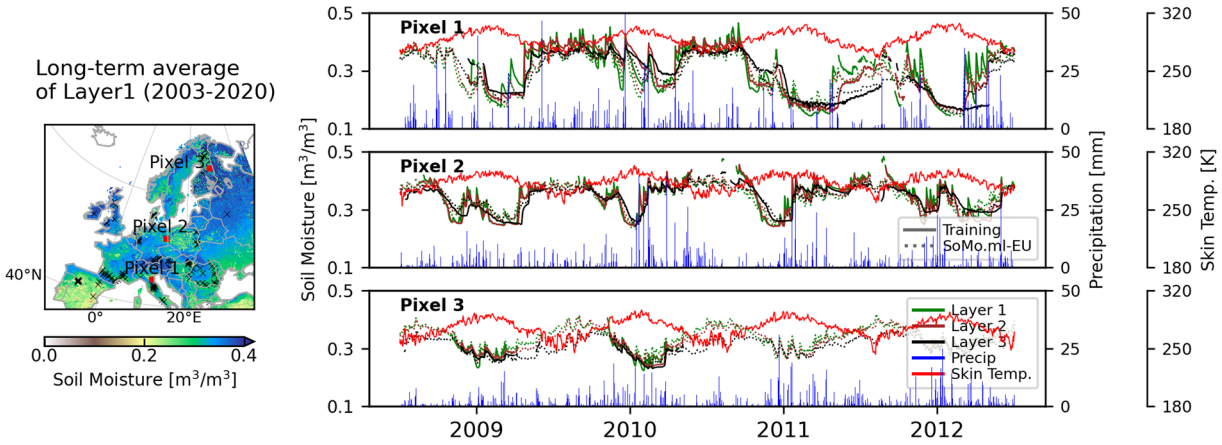

Fig. 7 Temporal behaviour of SoMo.ml-EU. Time series of soil moisture between the training data and SoMo.ml-EU (solid and dotted lines, respectively) at three selected pixels, where training soil moisture data are available for all three layers, are compared for the period from 2009 to 2012. Time series of precipitation and skin temperature inputs are also shown. All data are shown in 7-daily averages.

Usage Notes

Despite the availability of large-scale soil moisture datasets from diverse sources, soil moisture information is mostly provided at coarse spatial resolutions (25–50 km), which limits the consideration of land surface heterogeneity and corresponding physical processes related to, for example, land-climate coupling. To meet the increasing requirement for high-resolution data[30,33], we apply our distinctive approach, using ML trained with *in-situ* data, and deliver the high-resolution soil moisture over Europe. Consequently, SoMo.ml-EU offers daily soil moisture data at the spatial resolution of 0.1° over the period 2003–2020.

During the training-validation process, we confirm that ML can efficiently transfer knowledge from gauged to ungauged locations, even across continents. In our case, this is particularly useful for the simulations over arid regions because, in Europe, most *in-situ* data come from moderate climatic regions. Therefore, observational data covering more extreme conditions (e.g. arid regions) is only available from US networks. On the other hand, it remains challenging to obtain reliable soil moisture data for cold and humid regions, which could potentially introduce relatively high uncertainty in the final dataset in these regions. Performance of ML is known to be relatively poor in the conditions or locations that are not covered during training;[40,57] therefore, collecting more diverse and representative training data is imperative to improve the model performance. Over several years, more and more *in-situ* networks have been participating in the ISMN, and this study also utilises the *in-situ* data from the new network COSMOS-Europe. While those new data could be used in model training, we decided to keep them as independent data for validation purposes. Nonetheless, it is worth paying attention to publicly available big data because more diverse observational data is the key to reliable ML performance.

With the growing Earth observation data, ML has become a powerful tool for modelling the Earth system processes and has been widely used as a data-driven model to predict various hydrologic variables such as runoff[58] or evapotranspiration fluxes[59]. When using our soil moisture data in combination with other hydrological datasets that ML generates, water balance may not close due to biases in individual datasets. This is because ML does not explicitly take into account the balance between input and output water of a hydrological system, potentially introducing biases in the resulting data. Recently, hybrid modelling that combines physically-based (e.g. mass conservation) and ML approaches deriving water variables jointly is proposed to address this limitation[60].

Our new dataset SoMo.ml-EU will benefit meteorological and hydrological applications that require higher-resolution soil moisture data, e.g. evaluation of up/downscaled products, analyses of the spatial variability of hydrological processes within catchments, investigation of regional drought events and the role of soil moisture for hydrological dynamics and land-climate interactions. The high temporal correlation between the SoMo.ml-EU and *in-situ* data is mainly due to the uniqueness of our approach in directly using *in-situ* measurements (i.e. daily anomalies) to inform the model. More importantly, we process the training data and operate the model at a higher spatial resolution than in the previous version, which better mirrors the spatial footprint of the *in-situ* measurements. Consequently, more efficient use is made of the information content of these measurements to the reproduce soil moisture dynamics over ungauged locations. The closer agreement with the *in-situ* data revealed in our specific evaluation does not necessarily mean that SoMo.ml-EU generally outperforms the other datasets, given the discrepancies among the datasets in terms of spatial scales or soil layer depths. Rather, it indicates the distinctive characteristic of our data. Consequently, SoMo.ml-EU can serve as a independent reference against which to evaluate other high-resolution datasets. At the same time thanks to its ML-based derivation it could be used jointly with other established datasets to form an ensemble of independent gridded state-of-the-art soil moisture datasets with potentially increased robustness in data sparse regions.

Code availability

The code for the LSTM-based soil simulations can be downloaded at https://github.com/osungmin/SciData2022_SoMo_EU.

Received: 25 April 2022; Accepted: 14 October 2022;
Published online: 14 November 2022

References

1. Seneviratne, S. I. *et al*. Investigating soil moisture–climate interactions in a changing climate: a review. *Earth Sci Rev.* **99**, 125–161, https://doi.org/10.1016/j.earscirev.2010.02.004 (2010).
2. Bolten, J. D., Crow, W. T., Zhan, X., Jackson, T. J. & Reynolds, C. A. Evaluating the utility of remotely sensed soil moisture retrievals for operational agricultural drought monitoring. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **3**, 57–66, https://doi.org/10.1109/JSTARS.2009.2037163 (2010).
3. Orth, R. & Seneviratne, S. I. Using soil moisture forecasts for sub-seasonal summer temperature predictions in Europe. *Clim Dyn.* **43**, 3403–3418, https://doi.org/10.1007/s00382-014-2112-x (2014).
4. Wanders, N., Karssenberg, D., de Roo, A., de Jong, S. M. & Bierkens, M. F. P. The suitability of remotely sensed soil moisture for improving operational flood forecasting. *Hydrol. Earth Syst. Sci.* **18**, 2343–2357, https://doi.org/10.5194/hess-18-2343-2014 (2014).
5. Martínez-Fernández, J., González-Zamora, A., Sánchez, N., Gumuzzio, A. & Herrero-JimÃ©nez, C. 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 (2016).
6. O, S., Hou, X. & Orth, R. Observational evidence of wildfire-promoting soil moisture anomalies. *Sci. Rep.* **10**, 11008, https://doi.org/10.1038/s41598-020-67530-4 (2020).
7. Kroll, J. *et al*. Spatially varying relevance of hydrometeorological hazards for vegetation productivity extremes. *Biogeosciences* **19**, 477–489, https://doi.org/10.5194/bg-19-477-2022 (2022).
8. Famiglietti, J. S., Ryu, D., Berg, A. A., Rodell, M. & Jackson, T. J. Field observations of soil moisture variability across scales: Soil moisture variability across scales. *Water Resour. Res.* **44**, https://doi.org/10.1029/2006WR005804 (2008).
9. Brocca, L., Ciabatta, L., Massari, C., Camici, S. & Tarpanelli, A. Soil moisture for hydrological applications: Open questions and new opportunities. *Water* **9**, 140, https://doi.org/10.3390/w9020140 (2017).
10. Rodell, M. *et al*. The global land data assimilation system. *Bull. Amer. Meteor.* **85**, 381–394, https://doi.org/10.1175/BAMS-85-3-381 (2004).
11. Naz, B. S., Kollet, S., Franssen, H.-J. H., Montzka, C. & Kurtz, W. A 3 km spatially and temporally consistent European daily soil moisture reanalysis from 2000 to 2015. *Sci. Data* **7**, 111, https://doi.org/10.1038/s41597-020-0450-6 (2020).
12. Muñoz-Sabater, J. *et al*. ERA5-land: a state-of-the-art global reanalysis dataset for land applications. *Earth Syst. Sci. Data* **13**, 4349–4383, https://doi.org/10.5194/essd-13-4349-2021 (2021).
13. Koster, R. D. *et al*. On the nature of soil moisture in land surface models. *J. Clim.* **22**, 4322–4335, https://doi.org/10.1175/2009JCLI2832.1 (2009).
14. Petropoulos, G. P., Ireland, G. & Barrett, B. Surface soil moisture retrievals from remote sensing: Current status, products & future trends. *Phys. Chem. Earth.* **83-84**, 36–56, https://doi.org/10.1016/j.pce.2015.02.009 (2015).

15. Dorigo, W. *et al*. 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 (2017).
16. Chan, S. *et al*. Development and assessment of the SMAP enhanced passive soil moisture product. *Remote Sens. Environ.* **204**, 931–941, https://doi.org/10.1016/j.rse.2017.08.025 (2018).
17. Yao, P. *et al*. A long term global daily soil moisture dataset derived from AMSR-E and AMSR2 (2002–2019). *Sci. Data* **8**, 143, https://doi.org/10.1038/s41597-021-00925-8 (2021).
18. Park, S. *et al*. Downscaling GLDAS soil moisture data in East Asia through fusion of multi-sensors by optimizing modified regression trees. *Water* **9**, 332, https://doi.org/10.3390/w9050332 (2017).
19. Mao, H., Kathuria, D., Duffield, N. & Mohanty, B. P. Gap filling of high–resolution soil moisture for SMAP/sentinel–1: A two–layer machine learning–based framework. *Water Resour. Res.* **55**, 6986–7009, https://doi.org/10.1029/2019WR024902 (2019).
20. Guevara, M., Taufer, M. & Vargas, R. Gap-free global annual soil moisture: 15 km grids for 1991–2018. *Earth Syst. Sci. Data* **13**, 1711–1735, https://doi.org/10.5194/essd-13-1711-2021 (2021).
21. O, S. & Orth, R. Global soil moisture data derived through machine learning trained with *in-situ* measurements. *Sci. Data* **8**, 170, https://doi.org/10.1038/s41597-021-00964-1 (2021).
22. Guo, Z., Dirmeyer, P. A., Gao, X. & Zhao, M. Improving the quality of simulated soil moisture with a multi-model ensemble approach. *Q.J.R. Meteorol. Soc.* **133**, 731–747, https://doi.org/10.1002/qj.48 (2007).
23. Bai, W. *et al*. The performance of multiple model-simulated soil moisture datasets relative to ECV satellite data in China. *Water* **10**, 1384, https://doi.org/10.3390/w10101384 (2018).
24. Reichstein, M. *et al*. Deep learning and process understanding for data-driven Earth system science. *Nature* **566**, 195–204, https://doi.org/10.1038/s41586-019-0912-1 (2019).
25. Geer, A. J. Learning earth system models from observations: machine learning or data assimilation. *Phil. Trans. R. Soc. A.* **379**, 20200089, https://doi.org/10.1098/rsta.2020.0089 (2021).
26. Zhang, Y., Keenan, T. F. & Zhou, S. Exacerbated drought impacts on global ecosystems due to structural overshoot. *Nat. Ecol. Evol.* **5**, 1490–1498, https://doi.org/10.1038/s41559-021-01551-8 (2021).
27. Bastos, A. *et al*. Vulnerability of European ecosystems to two compound dry and hot summers in 2018 and 2019. *Earth Syst. Dynam.* **12**, 1015–1035, https://doi.org/10.5194/esd-12-1015-2021 (2021).
28. O, S. *et al*. The role of climate and vegetation in regulating drought-heat extremes. *J. Clim*. 35, 5677–5685, https://doi.org/10.1175/JCLI-D-21-0675.1 (2022).
29. Meng, X., Wang, H., Chen, J., Yang, M. & Pan, Z. High-resolution simulation and validation of soil moisture in the arid region of Northwest China. *Sci. Rep.* **9**, 17227, https://doi.org/10.1038/s41598-019-52923-x (2019).
30. Peng, J. *et al*. A roadmap for high-resolution satellite soil moisture applications - confronting product characteristics with user requirements. *Remote Sens. Environ.* **252**, 112162, https://doi.org/10.1016/j.rse.2020.112162 (2021).
31. Sabaghy, S., Walker, J. P., Renzullo, L. J. & Jackson, T. J. Spatially enhanced passive microwave derived soil moisture: Capabilities and opportunities. *Remote Sens. Environ.* **209**, 551–580, https://doi.org/10.1016/j.rse.2018.02.065 (2018).
32. Nayak, H. P. *et al*. High-resolution gridded soil moisture and soil temperature datasets for the indian monsoon region. *Sci. Data* **5**, 180264, https://doi.org/10.1038/sdata.2018.264 (2018).
33. Vergopolan, N. *et al*. Field-scale soil moisture bridges the spatial-scale gap between drought monitoring and agricultural yields. *Hydrol. Earth Syst. Sci.* **25**, 1827–1847, https://doi.org/10.5194/hess-25-1827-2021 (2021).
34. Abbaszadeh, P. *et al*. High-resolution SMAP satellite soil moisture product: Exploring the opportunities. *Bull. Amer. Meteor.* **102**, 309–315, https://doi.org/10.1175/BAMS-D-21-0016.1 (2021).
35. Hochreiter, S. & Schmidhuber, J. Long Short-term memory. *Neural Comput.* **9**, 1735–1780, https://doi.org/10.1162/neco.1997.9.8.1735 (1997).
36. Gao, P. *et al*. Modeling for the prediction of soil moisture in litchi orchard with deep long short-term memory. *Agriculture* **12**, 25, https://doi.org/10.3390/agriculture12010025 (2021).
37. Li, Q. *et al*. An attention-aware LSTM model for soil moisture and soil temperature prediction. *Geoderma* **409**, 115651, https://doi.org/10.1016/j.geoderma.2021.115651 (2022).
38. Dorigo, W. A. *et al*. 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 (2011).
39. Dorigo, W. *et al*. The International Soil Moisture Network: serving Earth system science for over a decade. *Hydrol. Earth Syst. Sci.* **25**, 5749–5804, https://doi.org/10.5194/hess-25-5749-2021 (2021).
40. O, S., Dutra, E. & Orth, R. Robustness of process-based versus data-driven modeling in changing climatic conditions. *J. Hydrometeorol.* **21**, 1929–1944, https://doi.org/10.1175/JHM-D-20-0072.1 (2020).
41. Beck, H. E. *et al*. Global-scale regionalization of hydrologic model parameters. *Water Resour. Res.* **52**, 3599–3622, https://doi.org/10.1002/2015WR018247 (2016).
42. Mittelbach, H. & Seneviratne, S. I. A new perspective on the spatio-temporal variability of soil moisture: temporal dynamics versus time-invariant contributions. *Hydrol. Earth Syst. Sci.* **16**, 2169–2179, https://doi.org/10.5194/hess-16-2169-2012 (2012).
43. Amante, C. ETOPO1 1 arc-minute global relief model: Procedures, data sources and analysis. *National Geophysical Data Center, NOAA* https://doi.org/10.7289/V5C8276M (2009).
44. Wieder, W. Regridded harmonized world soil database v1.2. *ORNL Distributed Active Archive Center* https://doi.org/10.3334/ORNLDAAC/1247 (2014).
45. Kratzert, F. *et al*. Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrol. Earth Syst. Sci.* **23**, 5089–5110, https://doi.org/10.5194/hess-23-5089-2019 (2019).
46. LeCun, Y. A., Bottou, L., Orr, G. B. & Müller, K.-R. Efficient BackProp. In *Neural networks: tricks of the trade, Second Edition*, 9–48, https://doi.org/10.1007/978-3-642-35289-8_3 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2012).
47. Gauch, M. *et al*. Rainfall–runoff prediction at multiple timescales with a single long short-term memory network. *Hydrol. Earth Syst. Sci.* **25**, 2045–2062, https://doi.org/10.5194/hess-25-2045-2021 (2021).
48. O, S., Orth, R., Weber, U. & Park, S. K. High-resolution european daily soil moisture derived with machine learning (2003–2020). *Figshare* https://doi.org/10.6084/m9.figshare.c.5957127 (2022).
49. Martens, B. *et al*. 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 (2017).
50. Bogena, H. R. *et al*. COSMOS-europe: A european network of cosmic-ray neutron soil moisture sensors. *Earth Syst. Sci. Data*. **14**, https://doi.org/10.5194/essd-14-1125-2022 (2022).
51. Koster, R. D. & Suarez, M. J. Soil moisture memory in climate models. *J. Hydrometeorol.* 2, 558–570, https://doi.org/10.1175/1525-7541(2001)0022.0.CO;2 (2001).
52. Orth, R., Koster, R. D. & Seneviratne, S. I. Inferring soil moisture memory from streamflow observations using a simple water balance model. *J. Hydrometeorol.* **14**, 1773–1790, https://doi.org/10.1175/JHM-D-12-099.1 (2013).
53. Wu, W. & Dickinson, R. E. Time scales of layered soil moisture memory in the context of land–atmosphere interaction. *J. Clim*. **17**, 2752–2764, 10.1175/1520-0442(2004)017<2752:TSOLSM>2.0.CO;2 (2004).
54. McColl, K. A. *et al*. The global distribution and dynamics of surface soil moisture. *Nature Geosci.* **10**, 100–104, https://doi.org/10.1038/ngeo2868 (2017).

55. Laaha, G. *et al*. The european 2015 drought from a hydrological perspective. *Hydrol. Earth Syst. Sci.* **21**, 3001–3024, https://doi.org/10.5194/hess-21-3001-2017 (2017).
56. Ionita, M. *et al*. The European 2015 drought from a climatological perspective. *Hydrol. Earth Syst. Sci.* **21**, 1397–1419, https://doi.org/10.5194/hess-21-1397-2017 (2017).
57. Meyer, H. & Pebesma, E. Predicting into unknown space? estimating the area of applicability of spatial prediction models. *Methods Ecol. Evol.* **12**, 1620–1633, https://doi.org/10.1111/2041-210X.13650 (2021).
58. Ghiggi, G., Humphrey, V., Seneviratne, S. I. & Gudmundsson, L. GRUN: an observation-based global gridded runoff dataset from 1902 to 2014. *Earth Syst. Sci. Data* **11**, 1655–1674, https://doi.org/10.5194/essd-11-1655-2019 (2019).
59. Jung, M. *et al*. The FLUXCOM ensemble of global land-atmosphere energy fluxes. *Sci. Data* **6**, 74, https://doi.org/10.1038/s41597-019-0076-8 (2019).
60. Kraft, B., Jung, M., Körner, M., Koirala, S. & Reichstein, M. Towards hybrid modeling of the global hydrological cycle. *Hydrol. Earth Syst. Sci.* **26**, 1579–1614, https://doi.org/10.5194/hess-26-1579-2022 (2022).
61. Dabrowska-Zielinska, K. *et al*. Soil moisture in the Biebrza wetlands retrieved from Sentinel-1 imagery. *Remote Sens.* **10**, https://doi.org/10.3390/rs10121979 (2018).
62. Brocca, L. *et al*. Soil moisture estimation through ASCAT and AMSR-E sensors: An intercomparison and validation study across Europe. *Remote Sens. Environ.* **115**, 3390–3408, https://doi.org/10.1016/j.rse.2011.08.003 (2011).
63. Zreda, M. *et al*. COSMOS: the COsmic-ray Soil Moisture Observing System. *Hydrol Earth Syst Sci.* **16**, 4079–4099, https://doi.org/10.5194/hess-16-4079-2012 (2012).
64. Ikonen, J. *et al*. The Sodankylä *in situ* soil moisture observation network: an example application of ESA CCI soil moisture product evaluation. *Geosci. Instrum. Methods Data Syst.* **5**, 95–108, https://doi.org/10.5194/gi-5-95-2016 (2016).
65. Al-Yaari, A. *et al*. The AQUI soil moisture network for satellite microwave remote sensing validation in south-western France. *Remote Sens.* **10**, https://doi.org/10.3390/rs10111839 (2018).
66. Cobley, A., Hemment, D., Rowan, J., Taylor, N. & Woods, M. GROW soil moisture data. *University of Dundee* https://doi.org/10.15132/10000156 (2020).
67. Bircher, S., Skou, N., Jensen, K. H., Walker, J. P. & Rasmussen, L. A soil moisture and temperature network for SMOS validation in Western Denmark. *Hydrol. Earth Syst. Sci.* **16**, 1445–1463, https://doi.org/10.5194/hess-16-1445-2012 (2012).
68. Morbidelli, R., Saltalippi, C., Flammini, A., Rossi, E. & Corradini, C. Soil water content vertical profiles under natural conditions: matching of experiments and simulations by a conceptual model. *Hydrol. Process.* **28**, 4732–4742, https://doi.org/10.1002/hyp.9973 (2014).
69. Biddoccu, M., Ferraris, S., Opsi, F. & Cavallo, E. Long-term monitoring of soil management effects on runoff and soil erosion in sloping vineyards in Alto Monferrato (North–West Italy. *Soil Tillage Res.* **155**, 176–189, https://doi.org/10.1016/j.still.2015.07.005 (2016).
70. Beyrich, F. & Adam, W. Site and Data Report for the Lindenberg Reference Site in CEOP - Phase 1. *Berichte des Deutschen Wetterdienstes* (2007).
71. Sanchez, N., Martinez-Fernandez, J., Scaini, A. & Perez-Gutierrez, C. Validation of the SMOS L2 soil moisture data in the REMEDHUS network (Spain. *IEEE Trans. Geosci. Remote Sens.* **50**, 1602–1611, https://doi.org/10.1109/TGRS.2012.2186971 (2012).
72. Schaefer, G. L., Cosh, M. H. & Jackson, T. J. The USDA natural resources conservation service soil climate analysis network (SCAN. *J. Atmos. Ocean. Technol.* **24**, 2073–2077, https://doi.org/10.1175/2007JTECHA930.1 (2007).
73. Calvet, J.-C. *et al*. *In situ* soil moisture observations for the CAL/VAL of SMOS: the SMOSMANIA network. In *2007 IEEE International Geoscience and Remote Sensing Symposium*, 1196–1199, https://doi.org/10.1109/IGARSS.2007.4423019 (IEEE, 2007).
74. Marczewski, W. *et al*. Strategies for validating and directions for employing SMOS data, in the Cal-Val project SWEX (3275) for wetlands. *Hydrol. Earth Syst. Sci. Discuss.* **7**, 7007–7057, https://doi.org/10.5194/hessd-7-7007-2010 (2010).
75. Zacharias, S. *et al*. A network of terrestrial environmental observatories in Germany. *Vadose Zone J.* **10**, 955–973, https://doi.org/10.2136/vzj2010.0139 (2011).
76. Schlenz, F., dall'Amico, J. T., Loew, A. & Mauser, W. Uncertainty assessment of the SMOS validation in the upper Danube catchment. *IEEE Trans. Geosci. Remote Sens.* **50**, 1517–1529, https://doi.org/10.1109/TGRS.2011.2171694 (2012).
77. Bell, J. E. *et al*. U.S. Climate Reference Network soil moisture and temperature observations. *J. Hydrometeor.* **14**, 977–988, https://doi.org/10.1175/JHM-D-12-0146.1 (2013).
78. Kirchengast, G., Kabas, T., Leuprecht, A., Bichler, C. & Truhetz, H. WegenerNet: a pioneering high-resolution network for monitoring weather and climate. *Bull. Amer. Meteor.* **95**, 227–242, https://doi.org/10.1175/BAMS-D-11-00161.1 (2014).
79. Petropoulos, G. P. & McCalmont, J. P. An operational *in situ* soil moisture & soil temperature monitoring network for West Wales, UK: The WSMN network. *Sensors* **95**, 227–242, https://doi.org/10.3390/s17071481 (2014).
80. Beck, H. E. *et al*. MSWX: Global 3-hourly 0.1° bias-corrected meteorological data including near-real-time updates and forecast ensembles. *Bull. Amer. Meteor.* **103**, E710–E732, https://doi.org/10.1175/BAMS-D-21-0145.1 (2022).
81. Beck, H. E. *et al*. MSWEP V2 global 3-hourly 0.1° precipitation: Methodology and quantitative assessment. *Bull. Amer. Meteor.* **100**, 473–500, https://doi.org/10.1175/BAMS-D-17-0138.1 (2019).
82. Fischer, G. *et al*. Global agro-ecological zones assessment for agriculture (gaez 2008). *IIASA, Laxenburg, Austria and FAO, Rome, Italy* (2008).

Acknowledgements

This study is supported by Brain Pool program funded by the Ministry of Science and ICT through the National Research Foundation of Korea (NRF-2021H1D3A2A02040136). Rene Orth acknowledges funding from the German Research Foundation (Emmy Noether grant 391059971). We thank the ISMN (https://ismn.geo.tuwien.ac.at/en/networks/) and COSMOS-Europe (https://doi.org/10.34731/x9s3-kr48) participating networks for sharing their *in-situ* soil moisture data.

Author contributions

S.O. and R.O. designed the study. S.O. conducted the experiments. U.W. contributed to collecting the data. All authors contributed to the analysis of the results and to the writing of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-022-01785-6.

Correspondence and requests for materials should be addressed to S.O. or S.K.P.