Multistep Forecasting of Soil Moisture Using Spatiotemporal Deep Encoder–Decoder Networks

LU LI,^a YONGJIU DAI,^a WEI SHANGGUAN,^a NAN WEI,^a ZHONGWANG WEI,^a AND SURYA GUPTA^b

^a Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), Guangdong Province Key Laboratory for Climate Change and Natural Disaster Studies, School of Atmospheric Sciences, Sun Yat-sen University, Guangzhou, Guangdong, China

^b Soil and Terrestrial Environmental Physics, Department of Environmental Systems Science, ETH Zurich, Zurich, Switzerland

(Manuscript received 2 July 2021, in final form 14 December 2021)

ABSTRACT: Accurate spatiotemporal predictions of surface soil moisture (SM) are important for many critical applications. Machine learning models provide a powerful method for building an accurate and reliable predictive model of SM. However, the models used in recent studies have some limitations, including lack of spatial autocorrelation (SAC), vague representation of important features, and primarily focused on the one-step forecast. Thus, we proposed an attention-based convolutional long short-term memory model (AttConvLSTM) for multistep forecasting. The model includes three layers, spatial compression, axial attention, and encoder–decoder prediction, which are used for compressing spatial information, feature extraction, and multistep prediction, respectively. The model was trained using surface SM from the Soil Moisture Active Passive L4 product at 18-km spatial resolution over the United States. The results show that AttConvLSTM predicts 24 h ahead SM with mean R^2 and RMSE is equal to 0.82 and 0.02, respectively. Compared with LSTM, AttConvLSTM improves the model performance over 73.6% of regions, with an improvement of 8.4% and 17.4% in R^2 and RMSE, respectively. The performance of the model is mainly influenced by temporal autocorrelation (TAC). Moreover, we also highlight the importance of SAC on model performance, especially over regions with high SAC and low TAC. Our model is also competent for SM predictions from several hours to several days, which could be a useful tool for predicting all meteorological variables and forecasting extremes.

KEYWORDS: Soil moisture; Forecasting; Deep learning

1. Introduction

Surface soil moisture (SM) has a fundamental role in climate, hydrological, and ecological systems, as it acts to balance the interaction of heat energy and hydrological processes (Seneviratne et al. 2010). It is a crucial variable for the partitioning of precipitation (P) into infiltration and runoff (Crow et al. 2018). Thus, better predictions of surface SM could improve many critical applications such as drought monitoring and water management (Zhang et al. 2021).

Physical models are commonly employed to forecast SM. However, Brooks et al. (2015) indicated that complex hydrologic processes have not been accurately represented in physical models. In addition, machine learning (ML) models provide a powerful method for SM predictions, which could avoid describing the hydrologic processes. Many ML models have achieved satisfactory results for SM prediction including traditional ML models and deep learning (DL) models (Zaman and McKee 2014; Fang and Shen 2020).

For traditional ML models, support vector machine regression (SVR) is one of the often-used techniques for SM prediction. Zaman and McKee (2014) proposed a multivariate SVR model to forecast SM based on lagged SM, soil temperature (ST), and P and obtained $0.96R^2$. Tree-based models are also commonly used for SM prediction and achieve relatively good performance. Zeng et al. (2019) used random forest (RF) to estimate daily SM at multiple depths (5, 25, and 60 cm) based on satellite imagery and ground-measured SM data in Oklahoma and achieved high accuracy at different soil depths. An artificial neural network (ANN) is also widely used to forecast surface SM. Kolassa et al. (2018) applied an ANN to estimate global surface SM based on brightness temperatures and meteorological forcing data from the NASA Goddard Earth Observing System Model version 5 (GEOS-5) land modeling system. The result indicated that ANN has a significantly higher forecast skill than the GEOS-5 model when evaluated against in situ SM measurements. However, traditional ML models are difficult to use for big data and have limited generalization capability, which are difficult to adequately predict SM at a large spatial scale.

DL models are part of a broader family of ML models based on ANN. Long short-term memory (LSTM) is an effective DL model that is widely utilized to predict SM (Fang and Shen 2020). Fang et al. (2017) used a LSTM model to predict surface SM based on climatic forcing, soil texture, and terrain slope and discovered that LSTM outperformed traditional ML models (e.g., ANN). Fang and Shen (2020) further found that the

Denotes content that is immediately available upon publication as open access.

Supplemental information related to this paper is available at the Journals Online website: <https://doi.org/10.1175/JHM-D-21-0131.s1>.

Corresponding author: Yongjiu Dai, daiyj6@mail.sysu.edu.cn

Publisher's Note: This article was revised on 7 September 2022 to correct how the affiliations were listed for the first five authors.

DOI: 10.1175/JHM-D-21-0131.1

© 2022 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).

trained LSTM model could add value to the long-term predictions and capture long-term trends of surface SM. LSTM-based models could adequately learn temporal autocorrelation (TAC) for SM prediction. However, the spatial autocorrelation (SAC) of SM is highly dynamic and changes over time, which may be also useful for SM predictions (Chen et al. 2017; Pan et al. 2019).

Hu et al. (2018) utilized a convolutional neural network (CNN) to predict SM, which could satisfy the requirements of learning SAC by convolutional (Conv) operation. Although CNN captures the SAC of SM and provides reasonable performance for simple short-term prediction, they are unable to adequately learn TAC (like LSTM). Shi et al. (2015) proposed a convolutional long short-term memory (ConvLSTM) network for P nowcasting, which could both learn TAC by LSTM structure and involve SAC by Conv operation. The result indicated that ConvLSTM consistently outperformed both LSTM and state-of-the-art physical models for P prediction. Based on this, the outstanding performance of ConvLSTM on SM forecasting was also proven by Li et al. (2021) in comparison with LSTM and CNN.

In general, SM predictive ML models used in recent studies still have some disadvantages. First, these models are usually unaware of SAC. Second, DL models could automatically weigh certain features more than traditional ML models during the training processes, however, they could not sufficiently represent important features and may be easily influenced by some invalid noisy features (Liu et al. 2019). Third, previous studies usually focus on one-step forecasting rather than multistep forecasting, which limited the forecasting range.

Therefore, to address above mentioned challenges, we developed an axial attention ConvLSTM (AttConvLSTM) model for SM prediction. The contributions of this paper are primarily threefold: 1) We used ConvLSTM to learn both TAC and SAC of SM. 2) We developed an axial attention module to enhance the representation of important features (Liu et al. 2019). This module could condense valid features by giving them larger weights and reduce the impact of invalid features by giving them smaller weights. 3) We utilized the encoder–decoder structure to provide multistep forecasting, which is widely used by some powerful DL models such as the Transformer model (Vaswani et al. 2017).

This paper is outlined as follows. We introduce the dataset used in our study in section 2. AttConvLSTM is introduced and presented model training and tuning processes in section 3. In section 4, we train the Soil Moisture Active Passive (SMAP) SM over the contiguous United States (CONUS) by the model, and the results are shown in relation to the following four aspects: 1) the overall performance of AttConvLSTM, 2) the improvement of AttConvLSTM compared with traditional ML and state-of-the-art DL models, 3) critical factors that influence performance, and 4) the applicability of AttConvLSTM over different time scales. Concluding remarks are presented in section 5.

2. Data

SM participates in many complex physical processes and is greatly influenced by different factors. Recent studies (Li et al. 2021; Carranza et al. 2021) applied four types of variables to

predict SM: memory parameters (i.e., impact of SM memory), meteorological parameters (i.e., impact from external climate factors), land parameters (i.e., impact from land surface physical processes), and time variables (i.e., impact of seasonal, interannual and daily variation). We used the most important covariates for SM prediction instead of using all possible covariates in this study. The important covariates for forecasting SM were selected based on previous studies and our preliminary tests. The model was trained for SMAP L4 data using lagged SM, lagged P , and lagged ST (the lagged time is 10 time steps, i.e., 30 h), physiographic attributes (i.e., land cover and elevation), and time variables (i.e., month, day) as inputs. Precipitation and ST were extracted from the SMAP L4 data and physiographic attributes were obtained from the United States Geological Survey (USGS). SMAP globally measures surface SM (0–5 cm) using a passive L-band radiometer (Entekhabi et al. 2008) and it has been the most promising satellite for SM monitoring due to higher capacity for penetrating vegetation at the L band compared to that at the C or X band (Wigneron et al. 2018). SMAP L4 data are derived from SMAP brightness temperature (Reichle et al. 2017a) and provides a variety of geophysical fields at a 3-h time resolution on the global 9 km modeling grids. SMAP L4 SM data were validated against in situ observations, suggesting that this product substantially improves over the model-only estimates (Reichle et al. 2017b). The domain of this study was restricted to the CONUS (23°–53°N, 67°–150°W). Considering the computational efficiency, we upsampled the input features (i.e., P , SM and ST) from 9- to 18-km resolution by skipping every other pixel in spatial dimensions. Physiographic attributes were also spatially mapped in 18-km resolution. The time range of SMAP L4 data spanned from 31 March 2015 to 24 January 2019. Therefore, a total of 11 677 data points were obtained for each grid, which was enough to train DL models for predicting SM (Fang et al. 2017).

We provided an input of lagged 30 h (i.e., 10 time steps for SMAP L4) features for models to make a daily prediction (i.e., 8 time steps for SMAP L4). We separated the CONUS into different patches, and each patch contains 8×8 pixels. The spatial dimension of each pixel is 18 km (Fig. 1a). For AttConvLSTM, we used an input grid of $576 \text{ km} \times 576 \text{ km}$ (i.e., 32×32 patch) to predict a $144 \text{ km} \times 144 \text{ km}$ (i.e., 8×8 patch) target square located over center (Fig. 1a). For RF, SVR, LSTM, and ConvLSTM, we trained the model on each 8×8 patch.

3. Model

We proposed a DL SM prediction model (Fig. 1b) in this study. The model consists of the following three layers: spatial compressing layer, axial attention layer, and encoder–decoder ConvLSTM layer. These layers are used for compressing spatial information, feature extraction, and multistep prediction, respectively.

a. Spatial compressing layer

The aim of the spatial compression layer is to introduce more spatial information from different spatial scales with lower computing costs (Sønderby et al. 2020). The inputs are 32×32 pixels, and we used the spatial compression layer to

FIG. 1. (a) The study area and the training setting. (b) Diagram of the deep learning soil moisture (SM) prediction model. ConvLSTM means convolutional long short-term memory model. The shape of inputs and outputs of each layer is annotated in the figure.

upscale inputs to 8×8 pixels. The layer comprises three steps (Fig. 2):

- 1) We applied a series of simple operations (i.e., center cropping, pooling, and space to feature) to reduce the spatial dimension of the input from 32×32 to 16×16 for each patch (Figs. 2a and S1).

- 2) We added static characteristic information, i.e., elevation, latitude, longitude, plant cover, and time terms, to the feature generated from the previous step (Fig. 2b).
- 3) We used a vanilla CNN (a 3×3 convolution with 16 filters, a 2×2 max pooling layer with stride 2, the activation function is “ReLU”) to capture the spatial variability of

FIG. 2. Structure of spatial compressing layer. The number annotated is the dimension of the spatial dimension in our study.

FIG. 3. The structure of axial attention layers, including (a) feature attention layer, (b) spatial attention layer, and (c) temporal attention layer. The shape of inputs and outputs of each attention layer is shown.

the input patch (Fig. 2c) and reduced the spatial dimension of each patch from 16×16 to 8×8 .

b. Axial attention layers

An attention mechanism (Woo et al. 2018) is used for adaptive feature extraction for SM prediction. Based on this, we could enhance the representation power of important spatio-temporal features for SM prediction. Figure 3 depicts the process of three axial attention modules (i.e., feature, spatial and temporal). The corresponding calculation details are shown in appendix A. We applied these three axial attention modules to focus on “what,” “where,” and “when” of inputs are important for SM prediction.

The process of feature attention of each time step is shown in Fig. 3a. First, we used average and max pooling over the whole spatial dimension to compress the input feature map, which has been commonly adopted to compress spatial information (Zhou et al. 2016; Hu et al. 2018). Previous studies have empirically confirmed that exploiting average and max-pooled features greatly improves the representation compared with independently using each feature (Woo et al. 2018). Second, both pooling features generated from the previous step are forwarded to a vanilla NN (the number of cells in the hidden layer is 16) to produce our feature attention map. Third, we merged the output feature using element wise summation and sent it to a sigmoid function to generate weights for the features.

The process of spatial attention of each time step is shown in Fig. 3b. First, we applied average and max pooling along

the feature axis and concatenated them to generate an efficient feature descriptor. Second, we employed a vanilla CNN (3×3 convolution with 16 filters) to generate the spatial attention map from the concatenated pooling features and send it to a sigmoid function to generate the weight matrix of the spatial dimensions.

The process of temporal attention of each time step is shown in Fig. 3c. First, we aggregated the feature dimension by a CNN (1×1 convolution with 16 filters). Second, we flattened the spatial layer and apply a vanilla NN (16 hidden cells) to generate a temporal attention map. Then we employed a sigmoid function to generate weights for each time step. We concatenated attention layers which are sequenced by feature, spatial, and temporal attention. From this, we got the input features for the encoder–decoder forecast layer.

c. Encoder–decoder ConvLSTM layer

To predict SM in the long-term rather than only one-step-ahead forecasting, we utilized a ConvLSTM in encoder–decoder architecture to predict SM on the next day. ConvLSTM is calculated as follows (for further detail refer to Shi et al. 2015):

$$\text{Input gate: } \mathbf{I}_t = \sigma(\mathbf{W}_{xi} * \mathbf{X}_t + \mathbf{W}_{hi} * \mathbf{H}_{t-1} + \mathbf{W}_{ci} \circ \mathbf{C}_{t-1} + \mathbf{b}_i), \quad (1)$$

$$\text{Forget gate: } \mathbf{F}_t = \sigma(\mathbf{W}_{xf} * \mathbf{X}_t + \mathbf{W}_{hf} * \mathbf{H}_{t-1} + \mathbf{W}_{cf} \circ \mathbf{C}_{t-1} + \mathbf{b}_f), \quad (2)$$

FIG. 4. The structure of the encoder-decoder convolutional long short-term memory (ConvLSTM) layer.

$$\text{Output gate: } \mathbf{O}_t = \sigma(\mathbf{W}_{xo} * \mathbf{X}_t + \mathbf{W}_{ho} * \mathbf{H}_{t-1} + \mathbf{W}_{co} \odot \mathbf{C}_t + \mathbf{b}_o), \quad (3)$$

$$\text{Input node: } \mathbf{G}_t = \tanh(\mathbf{W}_{xc} * \mathbf{X}_t + \mathbf{W}_{hc} * \mathbf{H}_{t-1} + \mathbf{b}_c), \quad (4)$$

$$\text{Cell state: } \mathbf{C}_t = \mathbf{I}_t \odot \mathbf{G}_t + \mathbf{F}_t \odot \mathbf{C}_{t-1}, \quad (5)$$

$$\text{Hidden state: } \mathbf{H}_t = \mathbf{O}_t \odot \tanh(\mathbf{C}_t), \quad (6)$$

where t represents the t th step of the input time steps. The term $\mathbf{X}_t$ is the input, $\mathbf{C}_t$ is the cell state, and $\mathbf{H}_t$ is the hidden state. The asterisk denotes the Conv operator, and the circle denotes the dot product. The σ is the sigmoid function, and $\tanh$ represents the hyperbolic tangent function. The $\mathbf{W}_{x\sim}$ and $\mathbf{W}_{h\sim}$ are 2D Conv kernels for spatial dimension.

The architecture of the encoder-decoder layer is illustrated in Fig. 4. We use the ConvLSTM cell for the encoder and decoder. The encoding ConvLSTM compresses the whole input image pool into a hidden state tensor and the decoder unfolds this hidden state to give the final multistep prediction.

d. Model training

We used 60%, 20%, and 20% of the data for training, validation, and test, respectively, which is commonly applied for prediction tasks (Shi et al. 2015). This dataset is divided by time order to ensure that we do not involve training data in the future that may increase performance artificially. For a fair comparison in the performance of different traditional ML and DL models, we used the same model structure and hyperparameters that are used for DL models and tuned the hyperparameters (Table S1 in the online supplemental material) of each model on each patch.

FIG. 5. Performance of attention-based ConvLSTM model (AttConvLSTM). (a) The temporal mean of SM prediction value and (b) the difference between prediction and original SM value. (c) The root-mean-square error (RMSE) and (d) the coefficient of determination (R^2) of the model.

FIG. 6. Boxplot of coefficient of determination (R^2 , red box) and root-mean-square error (RMSE, blue box) of different machine learning and deep learning models including support vector machine regression (SVR), random forest (RF), long short-term memory (LSTM), ConvLSTM, and AttConvLSTM.

The detail of tuning process is shown in [appendix B](#). The backward propagation uses the Adam gradient descent method. For the loss function, we applied the mean square error (MSE) between observation and prediction.

e. Performance assessment criteria

The model accuracy is evaluated based on two widely employed classic error criteria, including the root-mean-square error (RMSE) and coefficient of determination (R^2). These criteria are calculated as

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{N}}, \quad (9)$$

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2}, \quad (10)$$

FIG. 7. The (a) spatial autocorrelation index and (b) temporal autocorrelation index of surface SM from Soil Moisture Active Passive (SMAP).

FIG. 8. Relationship between temporal autocorrelation (TAC) and spatial autocorrelation (SAC) with (a) R^2 and (b) RMSE. The x axis represents the deviation of TAC (i.e., TAC minus mean TAC), and the y axis represents the deviation of the SAC index. The colors of the dots denote (a) R^2 and (b) RMSE, respectively. Both panels include insets that represent the variation of R^2 in (a) and RMSE in (b) as TAC and SAC increased. A red asterisk denotes the mean value of SAC on the top line and TAC on the bottom line.

where y_i is the observation value of the i th time steps, $\hat{y}_i$ is the corresponding predicted value by our model, and $\bar{y}_i$ is the mean value of the observation value. Parameter N is the number of time steps, R^2 represents how much variation in the observation is explained by the model, and the RMSE represents the difference of observation and prediction.

f. Correlation criteria

As we previously mentioned, it is considered that SAC and TAC have strong spatial variability related to the different soil texture, precipitation, and temperature (Pan et al. 2019), which substantially influence the predicting ability. We used the local Moran index (MI) to represent the SAC, which is calculated as

$$MI_i = (X_i - \bar{X}) \sum_{j \neq i}^n \omega_{ij} (X_j - \bar{X}), \quad (11)$$

where i and j denote different grids. The term X is the SM, and $\bar{X}$ is the spatial mean of the whole CONUS; ω_{ij} is a weight matrix that is defined as 1 when i is adjacent to j ; otherwise, it is set to 0. Notably, we only used a commonly used Moran index which used the mean SM of CONUS as X in Eq. (11). However, the mean value could vary across space and could be different for different pixels. Thus, we also applied trend-surface analysis to calculate spatial trend on each pixel and got the corresponding Moran index (see supplemental material). The results present nearly the same spatial distribution with the commonly used Moran index.

We utilized the Pearson coefficient ($\hat{\gamma}$) to represent the TAC on each grid. It is calculated as

$$\hat{\gamma}(k) = \frac{\sum_{t=1}^{n-k} (Z_{t+k} - \bar{Z})(Z_t - \bar{Z})}{\sum_{t=1}^n (Z_t - \bar{Z})^2}, \quad (12)$$

where k is the lagged time step, Z_t is the value of the SM for time step t on the target grid, and $\bar{Z}$ is the average of all time steps on the target grid.

4. Results and discussion

a. Overall performance of AttConvLSTM

Figure 5 shows the average value of performance over 8 time steps of AttConvLSTM during the test period. The map of the prediction bias (difference in time average SM between prediction and observation values) shows the range of ± 0.02 (Fig. 5b), which reveals that the prediction value is highly consistent with the SMAP L4 data in the average state. Furthermore, the mean RMSE of the model is 0.021, which suggests that the proposed model could accurately predict SM in most regions of the United States (Fig. 5c). Despite the overall high fidelity of the forecasting product, there are regions with larger errors. The central CONUS generally has larger errors compared with other regions, especially the central south CONUS. The spatial pattern is consistent with the results of several previous studies that utilized different methods, such as estimating the loss process (Koster et al. 2017), and predictions by LSTM (Fang et al. 2017). Moreover, in the eastern regions, AttConvLSTM has small RMSE with

relatively larger mean values of SM. This finding suggests that AttConvLSTM could forecast better in humid and permafrost regions. The R^2 represents the ability to predict the variance of the observed SM. The proposed model achieves high performance ($R^2 > 0.75$) in most regions of the CONUS, especially over the western, southeastern, and south-central United States (Fig. 5d). The error map highlights the central northern CONUS, which is consistent with the larger error region on the map of the RMSE. Moreover, we observe sporadic low value pixels along the southern and eastern coasts.

In general, AttConvLSTM is capable of predicting SM with high accuracy (i.e., high R^2 and low RMSE) in most regions of the CONUS, particularly the western, southeastern, and northeastern United States. However, the proposed model has relatively poor ability over the central area, which is typically a climate transition zone, especially over the central north regions (i.e., relatively low R^2 and high RMSE). Notably, in central south regions, although there is a gap in the SM values, the forecast variance is consistent with observations (i.e., relatively high R^2 and low RMSE).

b. Improvements in AttConvLSTM

To show the improvement by AttConvLSTM, we compared AttConvLSTM with existing state-of-the-art models, including SVM, RF, and LSTM. Furthermore, we applied the original ConvLSTM to investigate the impact of axial attention operation. We did not consider a deep neural network (DNN) for comparison because we found that the performance of DNN was worse than LSTM, which is mainly caused by strong persistence of SM. Figure 6 depicts the performance for the CONUS of different models. The median RMSE value of five models is 0.030, 0.024, 0.034, 0.022, and 0.02, and the median R^2 is 0.467, 0.693, 0.751, 0.789, and 0.817. Generally, we could determine that AttConvLSTM, ConvLSTM, LSTM, and RF predict SM relatively well (i.e., median R^2 is larger than 0.6 and the median RMSE is less than 0.03). AttConvLSTM gives the best results for both indexes and has less outliers compared to other models. We also test a stricter metrics, i.e., mean squared error skill score (MSESS). The metric uses climatological mean as the reference forecast rather than average of all observation, which gives a more natural reference forecast compared to R^2 (see supplemental material for the detail calculation). Figure S2 shows the spatial distribution of MSESS over CONUS of different machine learning models (i.e., AttConvLSTM, ConvLSTM, LSTM, and RF). Moreover, when compared AttConvLSTM with existing state-of-the-art methods, AttConvLSTM is also substantially improved the MSESS for both spatial distribution and median state, which is consistent with the results of R^2 and RMSE.

When we individually consider the model improvements by different operations, first, when we are comparing LSTM with RF (green arrow), R^2 increases by approximately 8.4%, which is mainly attributed to the introduction of the memory transfer mechanism (Fang and Shen 2020). However, the RMSE also increases which reveals that LSTM fares better at dryer regions while RF fares slightly better at moister regions.

FIG. 9. (a) Map of Köppen-Geiger climate regions over CONUS and (b) the estimation density curves R^2 in different climate regions. The inset of (b) is the estimation of density curves of R^2 in arid, temperate, and cold regions. The dashed red line, dotted green line, and solid blue line represents arid, temperate, and cold regions, respectively.

Second, R^2 increases by approximately 5.1% and the RMSE decreases by approximately 8.3% compared with the CONUS after introducing the Conv operation (cyan arrow), which emphasizes the importance of SAC for forecasting at a large spatial scale. Third, the median R^2 increases by approximately 3.3% and the median RMSE decreases by 9.1% (magenta arrow) after using axial attention. In general, memory transfer, Conv, and axial attention rectified errors substantially and improved the performance, suggesting that there is merit in pursuing spatiotemporal DL model for SM prediction.

c. Critical factors that influence prediction models

TAC and SAC are two critical factors that could influence the performance of SM predictive models (Prasad et al. 2018). Figure 7 shows the spatial distribution of SAC and TAC

TABLE 1. Mean value of different metrics of different climate regions, including soil moisture (SM), coefficient of determination (R^2) and root-mean-square error (RMSE) of the model, temporal autocorrelation (TAC), spatial autocorrelation index (SAC), and determinate coefficient between model performance (R^2) and TAC.

Climate regions	SM	R^2	RMSE	TAC	SAC	$R(R^2, \text{TAC})$
Arid, desert	0.010	0.793	0.017	0.897	0.047	0.692
Arid, steppe	0.152	0.795	0.022	0.906	0.018	0.661
Temperate, dry summer	0.159	0.918	0.017	0.965	0.022	0.802
Temperate, dry winter	0.157	0.845	0.023	0.921	0.011	0.723
Temperate, no dry season	0.261	0.824	0.025	0.927	0.024	0.839
Cold, dry summer	0.169	0.882	0.019	0.970	0.010	0.536
Cold, dry winter	0.188	0.629	0.023	0.828	0.006	0.763
Cold, no dry season	0.245	0.712	0.020	0.870	0.025	0.814

among the CONUS. The high-value regions predominantly encompass the southwestern and eastern regions for the SAC index, and the high-value regions for the TAC index primarily consist of the western and southeastern regions. Moreover, both indexes are relatively low in the central, north, and northeastern United States. Based on this, we further investigate the impact of these two critical factors on model performance from three aspects, i.e., overall, different climate regions, and different models.

1) IMPACT ON OVERALL PERFORMANCE

We investigate the relationship between the overall performance of AttConvLSTM over CONUS with the SAC and TAC index. Nearly all the large R^2 values (Fig. 8a) locate over the first and fourth quadrants (high TAC) and the low RMSE values (Fig. 8b) concentrate over the first and second quadrants (high SAC) near the y axis (i.e., high SAC and relatively high TAC). Moreover, R^2 is monotonically increasing for the TAC and R^2 is increasing as SAC increases (subfigure/inset in Fig. 8a). However, the increasing trend of R^2 disappears as the SAC increases over the mean value, which suggests that the SAC barely has an impact on R^2 . The RMSE shows a relatively no significant correlation with both TAC and SAC compared to R^2 (subfigure/inset in Fig. 8b). However, the obvious decreased signal appears over the regions with large SAC and relatively large TAC (pink line in subfigure Fig. 8b), which is consistent with the previous results.

Generally, both TAC and SAC significantly influence the performance of AttConvLSTM. To predict the variance in observations, TAC is a main factor and SAC has a limited impact on it. Thus, previous models that unaware of SAC also provided a relatively presentable performance (Pan et al. 2019; Fang and Shen 2020). However, SAC is also a main factor to enhance predictive power, especially over some regions with high SAC and low TAC. Therefore, we emphasize that only focusing on extract TAC is not enough to generate a robust and accurate model.

2) IMPACT ON DIFFERENT CLIMATE REGIONS PERFORMANCES

We investigated the performance over different climate regions according to the Köppen–Geiger climate index (Fig. 9a) (Beck et al. 2018). A summary of the performance

over eight climate regions appears in Table 1. All R^2 values over eight climate regions exceed 0.6, and the RMSE is less than 0.025. Moreover, the results exhibit small differences in the RMSE values, confirming that the AttConvLSTM is robust and applicable for different climate regions. Figure 9b shows the probability density of the R^2 of AttConvLSTM in different climate regions. Generally, the regions with high prediction accuracy are in temperate regions, follows by arid and cold regions (Fig. 9b). For temperate regions, we discover that the temperate dry summer regions (i.e., Mediterranean climate) achieve the best performance, follows by the temperate dry winter regions (i.e., monsoon climate), and temperate no dry seasons (i.e., temperate marine climate). All R^2 values over these regions are larger than the 75th percentiles of R^2 (0.88). For arid regions, we noted that the desert regions and steppe regions achieve relatively satisfactory performance (median of R^2). For cold regions, cold dry winter regions, and cold no dry season regions perform relatively poorly (less than 25th percentiles of R^2), showing the obvious long-tailed distributions. This result suggests that AttConvLSTM is

FIG. 10. Deviation of TAC and SAC in different four typical climate regions. The inset represents the correlation between R^2 with TAC and SAC over the corresponding four regions.

FIG. 11. The difference of R^2 between AttConvLSTM with (a) RF and (b) LSTM, and (c) between LSTM and RF. The shadow in (a) denotes the regions with high TAC (>75th percentiles TAC). The dots in (b) denote the regions with high SAC (>75th percentiles SAC).

relatively unstable in these climate conditions. Notably, there is a substantial difference between cold dry summer regions and other cold regions.

To investigate the relationship between two main factors and the performance of different climate regions, we depict the TAC and SAC over four typical climate regions (Fig. 10). We discover that the points of temperate regions are mostly located over the first quadrant (i.e., high SAC and high TAC), which explains the best performance of temperate dry summer regions (Fig. 9b). Furthermore, we discover that the points over arid and cold regions are divided by the x axis, which suggests that SAC is the main factor that influences the performance over arid regions rather than TAC. Moreover, the points of cold dry winter regions and cold dry summer regions are divided by the y axis, revealing that the significant difference in performance between these two regions is caused by different TAC. Moreover, the correlation between performance and TAC mostly exceeds 0.6 (Fig. 10), confirming the substantial impact of TAC on performance. Notably, the arid steppe region has a similar TAC index with the arid desert regions but with a much smaller SAC. However, the performances of these two regions are similar, indicating that TAC controls the performance over relatively high TAC regions. Generally, the performance for most climate conditions is significantly influenced by the TAC. However, the impact of the SAC could not be disregarded, especially over regions with low TAC and high SAC.

3) IMPACT ON DIFFERENT MODELS PERFORMANCES

Figure 11 shows the difference in R^2 between AttConvLSTM and RF (Fig. 11a) and LSTM (Fig. 11b) models. Generally, the mean value of the difference in R^2 for the CONUS is 0.095 and 0.062, which suggests that AttConvLSTM almost gives a better performance for the whole CONUS than LSTM

and RF model. However, for regions with relatively high TAC (the hatched region in Fig. 11a), AttConvLSTM only improves less than 0.04 of the R^2 value, which indicates that RF and LSTM already give relatively good performance in these regions. We also discover that the improvement of LSTM compared to RF is mostly located over some regions with relatively low TAC (central and northeastern of United States) (Fig. 11c). Although our model performs relatively poor over some regions with high SAC (dotted regions in Fig. 11b) and relatively low TAC (e.g., southwestern and central United States), AttConvLSTM already significantly improves the performance when compared with RF and LSTM, further reinforcing the importance of SAC. Furthermore, we also estimate the difference of R^2 between AttConvLSTM and ConvLSTM (figure omitted). Although axial attention significantly improves the forecast skill, we notice that the improvement of this mechanism is evenly distributed in space rather than concentrated over some specific regions. We also observe that this improvement over spatial is not related to SAC and TAC. These findings further reinforce the previous results.

Notably, the RMSE varies spatially and temporally with the mean and variance of SM. Thus, this variability may in turn be related to the TAC and SAC to some extent. On the other hand, R^2 gives credit to forecasts for simply reproducing climatology while the climatology is also related to TAC. Therefore, it is noted that the impact of TAC and SAC on model performance may partially because TAC and SAC reflect (or measure) some intrinsic characteristics of SM (i.e., climatology, spatial and temporal variability of SM).

d. Applicability of forecast on different time scales

Environmental factors could affect SM on different temporal scales (Li et al. 2020). For example, P could be caused by a

FIG. 12. The difference of R^2 between the 3-h case and the (a) 6- and (b) 12-h case. The curves show the spatial mean soil moisture value for SMAP (black line) and forecast (pink line) in the 6- and 12-h cases. The red triangle shows the catastrophe point in the time series of soil moisture. Black dots indicate the regions with largest difference of temporal autocorrelation (>75 quartile). Catastrophic points represent the period between different seasons.

cyclone for a short time scale and a monsoon for a long time scale (Silvestri and Vera 2003; Bombardi et al. 2014), which may lead to different impacts on SM. Therefore, to investigate the applicability of AttConvLSTM on different time scales, we averaged the SMAP L4 data and further applied AttConvLSTM to predict eight time steps in advance for different time resolutions, i.e., 2-day SM for 6-h time resolution and 4-day SM for 12-h time resolution. The spatial average over CONUS for these experiments is shown in Fig. 12. The R^2 value of the 6-h case and 12-h case are 0.968 and 0.916, which is suggesting that our model could give a relatively satisfactory performance from several hours to several days in the mean state. Moreover, the performance decreases as the time scale increases, especially for the period between different seasons (e.g., 2018–2009). Figure 12 shows the difference in R^2 between the cases and the corresponding difference in the TAC. We discover that the map of the difference in R^2 between different time scales is concentrated in the central United States and corresponds to the different areas of TAC, which indicates a significant correlation. From this, we could figure that the decrease in performance is determined by the decrease in TAC.

5. Conclusions

We developed a DL model for SM prediction across time scales (i.e., use hourly data to make daily forecasts),

including a spatial compressing layer, axial attention layer, and encoder–decoder ConvLSTM layer. We used axial attention to ensure that our model focused on valid features and was unaware of invalid features. Furthermore, ConvLSTM inherits the advantage of learning TAC from LSTM and introduces SAC by Conv operation. Therefore, we utilized ConvLSTM in an encoder–decoder structure for multistep forecasting.

AttConvLSTM predicts SM with unprecedented accuracy with a mean RMSE of 0.02 and a mean R^2 of 0.817. Moreover, our model outperforms traditional ML models and existing state-of-the-art DL models. AttConvLSTM improves the model performance with an improvement of 8.4% and 17.4% for R^2 and RMSE, respectively, compared to LSTM. Furthermore, our model is also competent for SM prediction from several hours to several days, which provides a useful tool for long-term forecasting.

From three aspects of model performance, i.e., overall, different climate regions, and different models, we all discovered that TAC significantly influenced model performance over most regions. Furthermore, performance was also influenced by SAC, particularly over the regions with high SAC and relatively low TAC.

Acknowledgments. This work was supported by the Natural Science Foundation of China under Grants U1811464,

41730962, and 41805072, and the National Key R&D Program of China under Grant 2017YFA0604300.

APPENDIX A

Detail of Axial Attention

The shape of inputs ($\mathbf{X}$) is [batch size, input time steps, latitude, longitude, features]. Feature and spatial attention are applied on input for each time step ($\mathbf{X}_t$) and temporal attention is applied on the whole input $\mathbf{X}$.

a. Feature attention

Compress spatial dimension by pooling operations,

$$\mathbf{SM}_t, \mathbf{SA}_t = \text{SMpool}(\mathbf{X}_t), \text{SApool}(\mathbf{X}_t); \quad (\text{A1})$$

generate feature scores by ANN,

$$\tilde{\alpha}_t = \text{ReLU}[W(\mathbf{SM}_t, \mathbf{SA}_t) + \mathbf{b}]; \quad (\text{A2})$$

and generate weighted outputs,

$$\bar{\mathbf{X}}_t = \sigma(\tilde{\alpha}_t) * \mathbf{X}_t, \quad (\text{A3})$$

where Mpool means max pooling on spatial dimension. Apool means average pooling on spatial dimension. ReLU is rectified linear unit, and σ is the sigmoid function. The terms W and $\mathbf{b}$ are parameters of ANN. The attention score $\tilde{\alpha}_t$ represents the importance of each feature and $\bar{\mathbf{X}}_t$ is weighted output. In our study, the shape of $\mathbf{X}_t$ and $\bar{\mathbf{X}}_t$ is [8, 8, 16], the shape of $\mathbf{SM}_t, \mathbf{SA}_t$ is [1, 1, 16], the shape of $\tilde{\alpha}_t$ is [1, 16].

b. Spatial attention

Compress feature dimension by pooling operations,

$$\mathbf{FM}_t, \mathbf{FA}_t = \text{FMpool}(\mathbf{X}_t), \text{FApool}(\mathbf{X}_t); \quad (\text{A4})$$

generate feature scores by CNN,

$$\tilde{\beta}_t = \text{ReLU}[\mathbf{W}_s * (\mathbf{CM}_t; \mathbf{CA}_t) + \mathbf{b}_s]; \quad (\text{A5})$$

and generate weighted outputs,

$$\bar{\mathbf{X}}_t = \sigma(\tilde{\beta}_t) * \mathbf{X}_t, \quad (\text{A6})$$

where FMpool is max pooling on feature dimension, FApool is average pooling on feature dimension. The terms $\mathbf{W}_s$ and $\mathbf{b}_s$ are parameters of CNN. The asterisk means conv operation. The attention score $\tilde{\beta}_t$ represents the importance of each point over spatial dimension and $\bar{\mathbf{X}}_t$ is weighted output. In our study, the shape of $\mathbf{X}_t$ and $\bar{\mathbf{X}}_t$ is [8, 8, 16], the shape of $\mathbf{FM}_t, \mathbf{FA}_t$ is [8, 8], the shape of $\tilde{\beta}_t$ is [8, 8].

c. Temporal attention

Compress feature dimension by CNN with 1×1 Conv and flatten the spatial dimension,

$$\mathbf{c} = \text{flatten}(\mathbf{W}_{1 \times 1} * \mathbf{X} + \mathbf{b}). \quad (\text{A7})$$

generate temporal scores by ANN,

$$\tilde{\gamma} = \text{ReLU}(\mathbf{W}_T \mathbf{c} + \mathbf{b}_T). \quad (\text{A8})$$

and generate weighted outputs,

$$\bar{\mathbf{X}} = \sigma(\tilde{\gamma}) * \mathbf{X}, \quad (\text{A9})$$

where $\mathbf{W}_{1 \times 1}$ means 1×1 convolutional filter and $\mathbf{W}_T$ and $\mathbf{b}_T$ are parameters of ANN. The asterisk means convolutional operation. The attention score $\tilde{\gamma}$ represents the importance of each time step and $\bar{\mathbf{X}}$ is weighted output. In our study, the shape of $\mathbf{X}$ and $\bar{\mathbf{X}}$ is [10, 8, 8, 16], the shape of $\mathbf{c}$ is [10, 64], the shape of $\tilde{\gamma}$ is [10, 1].

APPENDIX B

Hyperparameters Tuning

Table S1 in the supplemental material shows the hyperparameters to be tuned and tuning methods, and we show the tuning processes of traditional ML models and DL models as follows.

a. Tune parameters of traditional ML models

We utilized the grid search method (a commonly used method for tuning hyperparameters of ML model) to tune the hyperparameters of SVR and RF. For each ML model, we trained different models with different hyperparameters settings using the training dataset, and then validated these models on the validation dataset. From this, we could select the “best” hyperparameters set for the ML model on every single grid, which could avoid the overfitting problems.

b. Tune parameters of DL models

For DL models (LSTM, ConvLSTM, AttConvLSTM), we tuned the important hyperparameters (see Table S1) using a two-stage tuning process. To fairly compare to predictive capabilities of different DL models, we used the same structure. Moreover, we all use the same simple three-layer structure for each DL model to save time and cost. Although we were aware that we did not get the best model for each grid or patch, we fairly compared the performance of DL models. We tuned and got the best LSTM on each patch, and used the same setting for ConvLSTM and AttConvLSTM.

First, for each patch, we tuned the model structure hyperparameters (i.e., the number of cells in hidden layers and batch size). We trained the LSTM model with different numbers of cells ([16, 32, 64]) and different batch sizes ([32, 64, 128]), and selected the best model with the minimal mean RMSE of models over the whole patch.

Second, we tuned the training hyperparameters (i.e., epoch, learning rate) on every single grid. Figure S7 shows the loss of training and validation dataset during training of 64 LSTM model on a patch. We found that loss does not change after the epoch larger than 40 (actually, the model is obviously overfitted when epoch is larger than 20), thus we set epoch as 40 and selected the best model (the

minimal loss on validates dataset) during the training process. We set the learning rate as 0.01 and tuned learning rate by scaling the learning rate (scale factor is 0.1) when the MSE of the model was not decreased.

Notably, the tuning process significantly increases the training time, thus it is hard to tune model structure hyperparameters on each patch. However, we gave the same model structure on each patch for DL models and tuned the training parameters on each patch. Therefore, we already performed a fair comparison of DL and traditional ML models.

REFERENCES

- Beck, H. E., N. E. Zimmermann, T. R. McVicar, N. Vergopolan, A. Berg, and E. F. Wood, 2018: Present and future Köppen-Geiger climate classification maps at 1-km resolution. *Sci. Data*, **5**, 180214, <https://doi.org/10.1038/sdata.2018.214>.
- Bombardi, R. J., L. M. V. Carvalho, C. Jones, and M. S. Reboita, 2014: Precipitation over eastern South America and the South Atlantic Sea surface temperature during neutral ENSO periods. *Climate Dyn.*, **42**, 1553–1568, <https://doi.org/10.1007/s00382-013-1832-7>.
- Brooks, P. D., J. Chorover, Y. Fan, S. E. Godsey, R. M. Maxwell, J. P. McNamara, and C. Tague, 2015: Hydrological partitioning in the critical zone: Recent advances and opportunities for developing transferable understanding of water cycle dynamics. *Water Resour. Res.*, **51**, 6973–6987, <https://doi.org/10.1002/2015WR017039>.
- Carranza, C., C. Nolet, M. Pezij, and M. van der Ploeg, 2021: Root zone soil moisture estimation with Random Forest. *J. Hydrol.*, **593**, 125840, <https://doi.org/10.1016/j.jhydrol.2020.125840>.
- Chen, H., L. Fan, W. Wu, and H. Bin Liu, 2017: Comparison of spatial interpolation methods for soil moisture and its application for monitoring drought. *Environ. Monit. Assess.*, **189**, 525, <https://doi.org/10.1007/s10661-017-6244-4>.
- Crow, W. T., F. Chen, R. H. Reichle, Y. Xia, and Q. Liu, 2018: Exploiting soil moisture, precipitation, and streamflow observations to evaluate soil moisture/runoff coupling in land surface models. *Geophys. Res. Lett.*, **45**, 4869–4878, <https://doi.org/10.1029/2018GL077193>.
- Entekhabi, D., E. Njoku, P. O. Neill, M. Spencer, T. Jackson, J. Entin, E. Im, and K. Kellogg, 2008: The Soil Moisture Active/Passive Mission (SMAP). *IGARSS 2008 – 2008 IEEE International Geoscience and Remote Sensing Symp.*, Boston, MA, Institute of Electrical and Electronics Engineers, 3–6, <https://doi.org/10.1109/IGARSS.2008.4779267>.
- Fang, K., and C. Shen, 2020: Near-real-time forecast of satellite-based soil moisture using long short-term memory with an adaptive data integration kernel. *J. Hydrometeorol.*, **21**, 399–413, <https://doi.org/10.1175/JHM-D-19-0169.1>.
- , —, D. Kifer, and X. Yang, 2017: Prolongation of SMAP to spatiotemporally seamless coverage of continental U.S. using a deep learning neural network. *Geophys. Res. Lett.*, **44**, 11 030–11 039, <https://doi.org/10.1002/2017GL075619>.
- Hu, Z., L. Xu, and B. Yu, 2018: Soil moisture retrieval using convolutional neural networks: Application to passive microwave remote sensing. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, **XLII-3**, 583–586, <https://doi.org/10.5194/isprs-archives-XLII-3-583-2018>.
- Kolassa, J., and Coauthors, 2018: Estimating surface soil moisture from SMAP observations using a Neural Network technique. *Remote Sens. Environ.*, **204**, 43–59, <https://doi.org/10.1016/j.rse.2017.10.045>.
- Koster, R. D., R. H. Reichle, and S. P. P. Mahanama, 2017: A data-driven approach for daily real-time estimates and forecasts of near-surface soil moisture. *J. Hydrometeorol.*, **18**, 837–843, <https://doi.org/10.1175/JHM-D-16-0285.1>.
- Li, L., and Coauthors, 2020: A causal inference model based on random forests to identify the effect of soil moisture on precipitation. *J. Hydrometeorol.*, **21**, 1115–1131, <https://doi.org/10.1175/JHM-D-19-0209.1>.
- Li, Q., Z. Wang, W. Shangguan, L. Li, Y. Yao, and F. Yu, 2021: Improved daily SMAP satellite soil moisture prediction over China using deep learning model with transfer learning. *J. Hydrol.*, **600**, 126698, <https://doi.org/10.1016/j.jhydrol.2021.126698>.
- Liu, Y., Q. Zhang, L. Song, and Y. Chen, 2019: Attention-based recurrent neural networks for accurate short-term and long-term dissolved oxygen prediction. *Comput. Electron. Agric.*, **165**, 104964, <https://doi.org/10.1016/j.compag.2019.104964>.
- Pan, J., W. Shangguan, L. Li, H. Yuan, S. Zhang, X. Lu, N. Wei, and Y. Dai, 2019: Using data-driven methods to explore the predictability of surface soil moisture with FLUXNET site data. *Hydrol. Processes*, **33**, 2978–2996, <https://doi.org/10.1002/hyp.13540>.
- Prasad, R., R. C. Deo, Y. Li, and T. Maraseni, 2018: Soil moisture forecasting by a hybrid machine learning technique: ELM integrated with ensemble empirical mode decomposition. *Geoderma*, **330**, 136–161, <https://doi.org/10.1016/j.geoderma.2018.05.035>.
- Reichle, R. H., and Coauthors, 2017a: Assessment of the SMAP Level-4 surface and root-zone soil moisture product using in situ measurements. *J. Hydrometeorol.*, **18**, 2621–2645, <https://doi.org/10.1175/JHM-D-17-0063.1>.
- , and Coauthors, 2017b: Global assessment of the SMAP Level-4 surface and root-zone soil moisture product using assimilation diagnostics. *J. Hydrometeorol.*, **18**, 3217–3237, <https://doi.org/10.1175/JHM-D-17-0130.1>.
- Seneviratne, S. I., T. Corti, E. L. Davin, M. Hirschi, E. B. Jaeger, I. Lehner, B. Orlowsky, and A. J. Teuling, 2010: 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>.
- Shi, X., Z. Chen, H. Wang, D. Y. Yeung, W. K. Wong, and W. C. Woo, 2015: Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, C. Cortes et al., Eds., Neural Information Processing Systems, 802–810.
- Silvestri, G. E., and C. S. Vera, 2003: Antarctic oscillation signal on precipitation anomalies over southeastern South America. *Geophys. Res. Lett.*, **30**, 2115, <https://doi.org/10.1029/2003GL018277>.
- Sønderby, C. K., and Coauthors, 2020: MetNet: A neural weather model for precipitation forecasting. arXiv, 17 pp., <https://arxiv.org/abs/2003.12140>.
- Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, 2017: Attention is all you need. *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, I. Guyan et al. Eds., Neural Information Processing Systems, 5999–6009.
- Wigneron, J., and Coauthors, 2018: SMOS-IC: Current status and overview of soil moisture and VOD applications. *IGARSS*

- 2018 – 2018 *IEEE International Geoscience and Remote Sensing Symp.*, Valencia, Spain, Institute of Electrical and Electronics Engineers, 1451–1453, <https://doi.org/10.1109/IGARSS.2018.8519382>.
- Woo, S., J. Park, J. Y. Lee, and I. S. Kweon, 2018: CBAM: Convolutional block attention module. *Computer Vision – ECCV 2018*, V. Ferrari et al., Eds., Lecture Notes in Computer Science, Vol. 11211, Springer, 3–19, https://doi.org/10.1007/978-3-030-01234-2_1.
- Zaman, B., and M. McKee, 2014: Spatio-temporal prediction of root zone soil moisture using multivariate relevance vector machines. *Open J. Mod. Hydrol.*, **4**, 80–90, <https://doi.org/10.4236/ojmh.2014.43007>.
- Zeng, L., S. Hu, D. Xiang, X. Zhang, D. Li, L. Li, and T. Zhang, 2019: Multilayer soil moisture mapping at a regional scale from multisource data via a machine learning method. *Remote Sens.*, **11**, 284, <https://doi.org/10.3390/rs11030284>.
- Zhang, R., and Coauthors, 2021: Assessment of agricultural drought using soil water deficit index based on ERA5-land soil moisture data in four southern provinces of China. *Agriculture*, **11** (5), 411, <https://doi.org/10.3390/agriculture11050411>.
- Zhou, B., A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, 2016: Learning deep features for discriminative localization. *Proc. IEEE Conf. on Computer Vision Pattern Recognition (CVPR)*, Las Vegas, NV, Institute of Electrical and Electronics Engineers, 2921–2929, <https://doi.org/10.1109/CVPR.2016.319>.