Spatiotemporal Prediction of Alpine Vegetation Dynamic Change Based on a ConvGRU Neural Network Model: A Case Study of the Upper Heihe River Basin in Northwest China

Lifeng Zhang, Haowen Yan, Yi He, Sheng Yao, Shengpeng Cao, and Qiang Sun

Abstract—Accurate and comprehensive vegetation prediction methods are essential for effective agricultural planning and budgeting. Most existing vegetation prediction methods rely on sampling points rather than on overall spatiotemporal characteristics, making it difficult to accurately forecast vegetation changes. Hence, we built a neural network model with encoding and decoding modules based on a convolution gate recurrent unit (ConvGRU) and applied it to spatiotemporal normalized difference vegetation index (NDVI) predictions for the upper Heihe river basin (UHRB) in the Qilian Mountains, China. Based on MODIS NDVI raster data for the UHRB for 2000–2020, the analysis of the region’s spatiotemporal characteristics showed that the NDVI varied significantly over time. To avoid the gradient disappearance problem during ConvGRU model prediction, we proposed several numerical scaling methods for preprocessing the data before conducting training and prediction. We then constructed a spatiotemporal prediction model based on the ConvGRU, trained and predicted the numerically scaled data, and used various metrics to evaluate the model. The results showed that the grouping tanh-in function fitting was the least erroneous, and the ConvGRU model using data scaled by this method performed well across various metrics according to the test-set results. Also, unlike traditional time series prediction methods, the model accounted for spatiotemporal correlation features, and the output data of the prediction model were continuous and intuitive. Therefore, the proposed method is suitable for predicting dynamic vegetation changes. The NDVI prediction trend analysis indicated that the vegetation in the UHRB should improve in the future.

Index Terms—Deep learning, Qilian mountains, spatiotemporal prediction, vegetation dynamic change.

V. Introduction

Vegetation is a primary producer and an important component of terrestrial ecosystems. It links soil, hydrology, atmosphere, and other ecological elements, and has an irreplaceable regulatory role in the global material and energy cycle [1]. In recent years, multiple vegetation changes have occurred due to dramatic climate change and anthropogenic disturbances, which may have long-term impacts on forest resources, the climate, and food production [2]. Moreover, vegetation significantly affects surface stability, radiative properties, hydrological properties, and the global carbon cycle and atmospheric components [3], [4]. Therefore, regular monitoring and prediction of vegetation indices are important to ensure vegetation stability, maintain sustainable food production, prevent socioeconomic losses, effectively guide regional ecological restoration and environmental management, and reflect ecosystem status and functioning [5], [6].

The rapid development of remote sensing technology has led to the easy acquisition of satellite image time series data. It has become the main means to quantitatively assess vegetation growth, biomass, and coverage, and has been widely used in vegetation change monitoring at different scales [7]. The normalized difference vegetation index (NDVI) closely reflects vegetation coverage, biomass, leaf area index, and so forth. It is the most commonly used academic indicator for characterising the status of surface vegetation [8]. Moreover, it is sensitive to vegetation growth and potential, can effectively identify dynamic changes in surface vegetation coverage, and is considered the best surface vegetation indicator [9]. This index is a powerful tool for studying the ecological environment and is commonly used for vegetation prediction and management [10]. It is also widely used to monitor and predict agricultural production and to assess ecological and environmental changes [11].

The Qilian mountains form the boundary between the Qinghai–Tibet Plateau and the arid and semiarid inland areas of Northwest China, providing an important ecological protection barrier for the Hexi Corridor [12]. The Heihe river basin represents a typical Qilian mountain watershed, the vegetation...
changes of which affect local socioeconomic factors and the ecological environment but also have an important impact on the ecological security of Western China [13]. The upper Heihe river basin (UHRB) is a typically cold, arid region with fragile vegetation ecology and little human disturbance in its high mountain areas [14]. The UHRB has complex, changeable terrain, with obvious water and heat gradients and a significant vertical zonal distribution of vegetation [15]. The spatiotemporal variation of vegetation coverage in the UHRB can be used to facilitate a clearer understanding of the Qilian Mountain ecosystem, which is highly significant for establishing scientific ecological restoration and compensation mechanisms. So far, research on vegetation in the UHRB has focused primarily on the spatiotemporal characteristics of vegetation change. For example, Gao et al. [16] showed that vegetation in the UHRB in 2002 exhibited overall degradation, and Lei et al., based on UHRB NDVI data from 2001 to 2014, found that the NDVI had improved overall in the past 14 years. Tan et al. [17] used UHRB NDVI data from 2001 to 2017 and found that vegetation in the UHRB had increased slightly over the past 17 years [18]. Ren et al. [14] studied variations in the spatiotemporal characteristics of vegetation in the UHRB from 2001 to 2017 and discovered that the interannual variation in the NDVI increasingly fluctuated. Using long-term NDVI time series data from 1990 to 2019, Wang et al. analysed dynamic changes in vegetation coverage in the UHRB. The results showed that vegetation coverage increased [19]. Previous studies have mainly focused on analysing past and current spatiotemporal vegetation changes in the UHRB, but variations in the accuracy of the data have affected the results, making them inconsistent and sometimes contradictory. It is believed that vegetation degradation in watersheds will increase by the middle of the 21st century [20], which has prompted an increasing need to monitor and predict spatiotemporal changes in the region’s vegetation.

Currently, most traditional prediction methods are based on mathematical–statistical modelling. Fernandez-Manso et al. [21] used the autoregressive integrated moving average statistical method to predict NDVI changes in Castile and León (Spain) than ten days. They confirmed that time series analysis can predict the NDVI in the short term. However, their model assumed that the NDVI time series were linear and stable, which failed to effectively capture the non-linear, non-stationary information hidden in the time series, resulting in poor NDVI prediction performance [22]. Moreover, this method, based on mathematical–statistical modelling, models each sampling point separately, meaning that the method is not suitable for large-scale NDVI predictions.

To deal with this problem, some scholars have used machine learning for NDVI predictions. Ma et al. [23] compared the effects of four machine-learning methods for NDVI predictions of karst landforms. The results showed that the support vector regression (SVR) method had good prediction accuracy. However, SVR is also based on single sampling points, which causes problems like those of mathematical–statistical models. The development of deep-learning technology has also allowed NDVI predictions based on recurrent neural network (RNN) models [24], [25], [26]. RNNs can model multiple sampling points to make predictions. However, traditional RNN weight calculations rely on full connection operations, resulting in serious space-related redundancy and huge numbers of parameters, which undermine efficiency and performance.

Although existing NDVI prediction methods typically use historical sampling point data, the neighbourhood relationships between temporal and spatial changes in the NDVI mean that historical sampling point data often lack spatial neighbourhood information. Spatiotemporal prediction methods for large ranges of NDVI raster data are more intuitive, and they simplify the further analysis of prediction trends.

Considering the above problems, we built a neural network model with a pre-encoding module and a postdecoding module to predict the NDVI for the complete Heihe River Basin. It is based on a convolution gate recurrent unit (ConvGRU) with both convolution and recurrent structures [27]. Using raster data for predictions can provide spatially continuous and regular NDVI predictions, making it easier to assess the overall growth of vegetation in the UHRB and to identify potential risks. Since the NDVI values in the study area changed significantly over a year, most of the pixel values were relatively small. To correct the error accumulation and gradient disappearance in the subsequent predictions of the ConvGRU model, we applied three fitting methods (Fourier series, grouping univariate linear, and grouping tanh-ln functions) to fit the time series NDVI raster extremes, and we conducted numerical scaling separately for subsequent comparisons of the prediction results. After using the ConvGRU model for spatiotemporal predictions of the NDVI for the complete UHRB, we selected the optimal scaling method by combining multiple evaluation metrics to verify the accuracy and reliability of the NDVI data predicted by the ConvGRU model. Finally, we analysed the predicted NDVI for the UHRB to assess vegetation growth and identify potential risks.

II. STUDY AREA AND DATA SOURCE

A. Overview of the Study Area

Located in the middle of the Hexi Corridor, the Heihe River Basin is the second largest inland river basin in the arid region of Northwest China. It is adjacent to the Maying River in the east and the Lai and Fengle Rivers in the west. The total length of the Heihe river basin’s main stream is 821 km. It originates in the centre of the northern foothills of the Qilian Mountains in Qinghai, flows through the Qinghai and Gansu provinces and the Inner Mongolia autonomous regions, and terminates in Juyanhai in the Ejina Banner. The total drainage area is about 14.29 × 10 km². The elevations in the UHRB decrease from south to north, with the lowest elevation at 2071 m and the highest at 4879 m (Fig. 1).

The basin’s geological conditions are complex, and the hydrometeorology, soil, and vegetation are unique. The upper, middle, and lower reaches of the basin have different ecological functions. The UHRB has plentiful ice and snow and relatively abundant rainfall, making the region an ecologically functional area for glaciers and water conservation [28]. In recent years, the Chinese government has implemented a series of ecological restoration measures in the UHRB, including grazing prohibition, grassland restoration, and forest protection in water
conservation areas. Therefore, vegetation coverage in the region is increasing.

B. Data Source and Preprocessing

The vegetation data used for this article were derived from MOD13A1 data provided by the United States National Aeronautics and Space Administration (NASA) for a time series of 2000–2020. The temporal resolution was 16 days (23 scenes per year), and the spatial resolution was 500 m. The data can be downloaded from the NASA website.¹ We used the MODIS Reprojection Tool to perform preliminary processing, such as splicing and projection conversion of the downloaded MODIS NDVI data. We then used ArcMap software to enter the minimum values for the neighborhood into the processed remote-sensing images to fill in gaps and address outliers, and then we cut them. To facilitate the design of the subsequent neural network model and ensure that the entire study area was included, we cropped the NDVI raster data into a square of 768×768 pixels.

We then synthesized the 16-day NDVI data into monthly data using the maximum value composite method to eliminate interference from clouds, the atmosphere, and solar altitude [29]. During the trend analysis, we further used the NDVI for each month to calculate the annual NDVI and to analyze the long-term NDVI variation characteristics.

We drew the digital elevation model (DEM) data used in this article from NASA’s SRTM data,² with a spatial resolution of 90 m. We resampled the 90 m DEM image to match the 500 m spatial resolution using the nearest neighbor sampling method in the ArcGIS software.

III. RESEARCH METHODS

A. Theil–Sen (TS) Median Trend and Mann–Kendall (MK) Test

We used a combined TS median trend and MK test to identify trends in the long-term time series data [30]. TS median trend analysis is a robust nonparametric statistical trend calculation method that can reliably estimate monotonic trends. The MK test is a non-parametric test, which, unlike parametric test methods, does not require samples to follow a certain distribution and is not disturbed by a few outliers; hence, it is more suitable for ordinal variables. The MK test has been successfully applied to hydrological and meteorological time series analysis [31], [32], and it can be used to assess increasing or decreasing trends in time series data. We used the TS–MK test to analyse annual NDVI variation trends for each raster from 2000 to 2020 and from 2020 to 2031. Positive MK values indicated an upward trend, and vice versa. Values of 1.645, 1.96, and 2.32 indicated that the MK trend levels were 10%, 5%, and 1%, respectively. NDVI trends were considered significant when the absolute value of MK was greater than 1.96. We also used R-squared ($R^2$) and $p$ values with a significance level of 0.05 to evaluate the effect of the single linear regression model [33].

B. Creating the Dataset

1) Numerical Scaling Method: In previous studies that used classical CNN and RNN models for classification or regression modelling, the numerical scale of the data was mostly processed to [0, 1] or [−1, 1] for training and prediction [34], [35], [36], [37]. The purpose of this process is to make it easy for the model to converge and to prevent the gradient from vanishing or exploding. Considering our experimental data, the range of the MODIS NDVI data was [−0.2, 1]. We tried to use the data for

¹[Online]. Available: https://ladsweb.modaps.eosdis.nasa.gov/
²[Online]. Available: http://srtm.csi.cgiar.org/
direct model training and prediction, but the overall NDVI value, particularly the peak value, in the study area changed greatly with the season. If this NDVI data had been used directly for model training and prediction, the numerical scale for the raster data in winter, early spring, and late autumn would have been too small, leading to a vanishing gradient in the prediction process for the future and a gradual error accumulation. Specifically, the overall value of the raster prediction decreased slowly over time. Compared with the historical trend, this phenomenon was obviously abnormal (as shown by the red dotted line in Fig. 11 in Section IV-C of the experimental results section).

To address this problem, we tried to scale the NDVI data using function fitting. We divided all pixel values of the raster data by the fitting value of the corresponding time step to stretch the maximum pixel value to near 1. After the model predicted the output result, it could be reversed (i.e., multiplied by the function-fitting value in the current time step).

We needed to use a function fitting process to determine the target sequence, optimisation algorithm, and function template. To reduce the impact of possible abnormal pixel values on the fitting process, we obtained the value of the current time step in the target sequence by calculating the average value of the maximum $K$ pixel values (TopK) in the corresponding raster data (where $K = 6$, equivalent to about 1/100 000 of the total number of pixels). We selected the Levenberg–Marquardt algorithm [38] as the optimisation algorithm because it is the most widely used nonlinear least squares algorithm for finding the parameter vector that minimises error in the fitting function. Over the course of the experiment, we gradually formulated three function templates and compared the fitting sequence errors for the three function templates (see Table I). Furthermore, after scaling the experimental data with the three function templates, we compared the prediction performance of the neural network models trained and predicted with the three processed datasets (see Table IV).

Due to annual seasonal variations, the target sequence changed periodically during the year (23 scenes) due to the target sequence comprising the average value of the maximum time series NDVI raster data. Therefore, we first tried to use a triangular expansion of the Fourier series as the basic function template

$$f_{\text{Furr}}(x) = a_0 + \sum_{n=1}^{N} \left[ a_n \cos \left( \frac{2\pi n}{T} x \right) + b_n \sin \left( \frac{2\pi n}{T} x \right) \right],$$

$$a_0 \in \mathbb{R} \tag{1}$$

where $a_0$, $a_1 \sim a_N$, and $b_1 \sim b_N$ are the parameters to be obtained; $N$ is the maximum order of the Fourier series (here $N = 8$); and $T$ is the period, aligned with the number of scenes per year, (here, $T = 23$). On this basis, to adapt to the interannual overall NDVI growth trend in the study area, we added a univariate linear function

$$g_{\text{fit}}(x) = k \cdot f_{\text{Furr}}(x) + b \tag{2}$$

where slope $k$ and intercept $b$ are the parameters to be solved. The final fitting results are shown in Fig. 2.

Fig. 2 shows that the fitting results for the Fourier series fitted most of the scatter points. Moreover, with increases in the independent variable (i.e., time step), the value of the fitting function also increased periodically. However, in subsequent predictions, the value of the fitting function gradually exceeded 1, which was unreasonable for the NDVI.

Considering that the periodic function template could not easily adapt to the details of the target sequence changing year by year, we decided to make the strategy of function fitting more flexible. We tried to group the maximum TopK mean of the time series NDVI raster data according to the scene number for each year to generate 23 target sequences, and we used the new function template to fit them. The univariate linear function is the most easily implemented function template. The formula $f(x) = kx + b$ means that this template only needs to obtain the parameter values for $k$ and $b$. Thus, because the function template needed to fit the 23 target sequences, it had to obtain 23 pairs of parameter values for $k$ and $b$. These fitting results are shown in Fig. 3.

As shown in Fig. 3, unlike using the Fourier series as a function template to fit the complete target sequence, group fitting can combine the fitting values of multiple univariate linear functions to obtain more flexible and complex linear fitting results. Compared with the Fourier series, this group fitting method not only produces a smaller fitting error but also significantly saves on computation overheads. However, since the slopes of univariate linear functions were constant (and for the study area, all positive), this led to the same problem as for the fitting method using the Fourier series. The values of the fitting function gradually exceeded 1 with time-step increases in predictions for the future. In particular, as Fig. 3(a) shows,
In summary, we needed to find a monotonically increasing function template with an upper bound. This function template had to make the value of the fitting function meet the growth situation, and its value range had to be less than 1. After many attempts, we found that the best fitting effect was obtained by using the natural logarithmic function (ln) to ensure monotonic increases, and we added the hyperbolic tangent function (\( \tanh \)) to limit the upper and lower bounds of the function template based on the univariate linear function. The function template formula was as follows:

\[
f(x) = \tanh[\ln(kx + b)]
\]  

(3)

where slope \( k \) and intercept \( b \) of the univariate linear function are the parameters to be solved. This function template was self-limiting. We did not add the parameters for other locations to avoid overall changes in the value range and other unpredictable situations. After fitting, the function template provided a fitting effect like that of the univariate linear function when the slope was small and the intercept was large. However, when the slope was large, it exhibited obvious bending and ensured that the upper bound of the function was 1. The fitting results using this function template are shown in Fig. 4.

Fig. 4 shows that the monotonically increasing and bounded function template completely avoided the problem of a fitting value exceeding 1. Moreover, the error of this fitting method was smaller than that of the previous two methods [see Table I, for which we used the root mean square error (RMSE) and mean absolute percentage error (MAPE)]. We also generated their datasets from the unprocessed and three kinds of processed data, used the model proposed below for training and prediction, respectively, and used a variety of evaluation methods for test-set comparisons (see Section IV-C and Table IV).

2) Creation of the Dataset: For this article, we adopted the year-to-year prediction method. Using NDVI raster data for 23 scenes per year as the input to the model, the model carried out spatiotemporal predictions for the study area’s NDVI values and their distribution for each scene in the next year. For all the time series data ranges, we used a sliding window to segment the
dataset by year. Due to the short duration of the original data in this article, we also used label data as the input, and then used the overlapping segmentation method to ensure a suitable amount of training data, as shown in Fig. 5.

Because MODIS data began in 2000, the data only covered 21 years by the end of our research. In the case of forecasting in years, the overall data-set capacity was small. After taking the data for 23 time steps in the last year of the sample as the label of the test set, we divided the remaining samples into a training set and a validation set by K-fold cross-validation (in this article, $K = 5$). This specific operation extracted the remaining data pairs and divided them into five parts in chronological order. Four parts (80%) were selected as the training set, and one part (20%) as the validation set. Finally, we saved them as five groups of training sets and validation sets. We trained the models and cross-verified the results of the test set.

C. Structure of the Network Model

In this article, we built an end-to-end neural network model based on a ConvGRU to predict time series NDVI raster data for the UHRB in the Qilian mountains. We divided the network structure into three main parts and used the encoder–predictor–decoder mode as the main framework to carry out the structural design. The first part was the feature extraction (encoding) module, the second part was the prediction module based on a ConvGRU, and the third part was the feature rebuilding output (decoding) module. The encoding module performed convolution and pooling operations on the time series NDVI raster data for feature extraction coding while reducing the size of the feature map, thus greatly reducing the computational overhead and storage space occupation of the subsequent ConvGRU prediction module. The prediction module used three bidirectional ConvGRU layers to perform a multistep spatiotemporal prediction of the time series feature map. The decoding module performed upsampling and detailed feature restoration decoding on the feature map output of the prediction module, and finally delivered the prediction results. The overall structure of the specific network model is shown in Fig. 6.

1) Encoding Module: The encoding module included four scales for feature extraction, using a time-distributed two-dimensional (2-D) convolution layer, a max pooling layer, an average pooling layer, and an instance normalisation layer [39]. The time-distributed 2-D convolution layer could accept a 5-D tensor, and the parameters of its convolution kernel were shared in each time step. This design replaced the 3-D convolution based on the premise of ensuring the feature extraction performance of the raster data to compress some parameters and maintain independence between each time step.

In the encoding module, the convolution kernel sizes of the three convolution layers located in the maximum (original) size feature map were $7 \times 7$, $1 \times 7$, and $7 \times 1$. All remaining
convolution layers had a convolution kernel size of $3 \times 3$. An identical feature map padding method was used for all the convolution layers to ensure that the output feature map was the same size as before convolution. The output channel of the convolution layer increased with decreases in the feature map size during the downsampling process. The output channels of the convolution layer located in the feature map had the largest size of 16, which gradually doubled to 32, 64, and 128. Their placements are shown in Fig. 6.

We reduced the size of the feature map from $768 \times 768$ in the initial model to $64 \times 64$ after three downsampling (i.e., pooling) operations with different factors. When the pooling operation was carried out with the maximum-sized feature map, the window size was $3 \times 3$ and the stride was 3. We applied both maximum and average pooling and concatenated the results. We chose this design to minimise the loss of feature details caused by large-scale downsampling equivalent to one-third of the current width and height. The other downsampling operations only used maximum pooling, with a window size of $2 \times 2$ and a stride of 2. The feature map padding methods for all pooling layers were valid, so we did not calculate the pixel value repeatedly, based on the premise of completely traversing all pixels. Thus, we successfully reduced the width and height of the feature map to the required size.

2) Prediction Module: The second part of the prediction module was based on a ConvGRU, which comprised three bidirectional, 2-D ConvGRU layers. The calculation process for a single ConvGRU cell is described by the following formula:

$$
\begin{align*}
    r_{t} &= \sigma(W_r \ast [h_{t-1}, x_{t}] + b_r) \\
    z_{t} &= \sigma(W_z \ast [h_{t-1}, x_{t}] + b_z) \\
    \hat{h}_{t} &= \tanh(W_h \ast [r_{t} \circ h_{t-1}, x_{t}] + b_h) \\
    h_{t} &= (1 - z_{t}) \circ h_{t-1} + (z_{t} \circ \hat{h}_{t})
\end{align*}
$$

where $W$ and $b$ are the weight matrix and the bias term of different calculation processes, respectively; $[h_{t-1}, x_{t}]$ represents the concatenation result of the output tensor of the previous time-step cell and the input tensor of the current cell; $\ast$ represents the convolution operation; $\circ$ represents the Hadamard product operation; $\sigma$ represents the sigmoid function; and $\tanh$ represents the hyperbolic tangent function. Fig. 7 shows the structure of a single ConvGRU cell.

The recurrent activation function (sigmoid) and the activation function ($\tanh$) in the ConvGRU layer were defaults, which were verified to meet the requirements and performance of this article.

The prediction module included three bidirectional ConvGRU layers. We used this design to capture the spatiotemporal correlation features in each time series direction for prediction, and to obtain more stable time series feature maps. We selected the concatenation method as the merge mode to operate in the time-distributed convolution layer and the upsampling layer, the feature map sizes were restored, and the feature details were gradually recovered. The kernel size of each convolution layer connected behind the upsampling layer was the same as for the upsampling factor to reduce the hard-edge problem caused by the nearest neighbour algorithm in the upsampling operation. The output channels of the convolution layers in the decoding module decreased as the size of the feature map increased. The front two upsampling layers expanded the size of the feature map from $64 \times 64$ to $256 \times 256$. The upsampling factor was 2 for both rows and columns, and the interpolation method was the nearest neighbour algorithm. The final upsampling operation had to expand the size of the feature map from $256 \times 256$ to $768 \times 768$ (i.e., a threefold expansion of the length and width). To make the detailed upsampling features as perfect as possible without introducing too many errors, we adopted two interpolation methods: the nearest neighbour algorithm and the bilinear algorithm. We concatenated the feature maps for the two sets of upsampling results in the channel dimension. We then carried out the final reconstruction and output NDVI raster data predictions using five convolution layers, the output channels of which were gradually reduced to 1.

D. Evaluation Methods

To quantitatively evaluate the performance of the spatiotemporal model predictions, we used three evaluation methods to evaluate the accuracy of the prediction results. The first evaluation assessed the numerical error of the model prediction results in practice, including the mean absolute error (MAE) and the
mean square error (MSE)

\[
\text{MAE} = \frac{1}{m} \sum_{i=1}^{m} |y_i - \hat{y}_i|
\]

\[
\text{MSE} = \frac{1}{m} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2.
\]

The second evaluation assessed the spatiotemporal regression correlation between the model prediction results and the labelled NDVI raster data, including the explained variance (EVar) and the coefficient of determination \(R^2\)

\[
\text{EVar} = 1 - \frac{\sum_{i=1}^{m} (y_i - \hat{y}_i - (\bar{y} - \bar{\hat{y}}))^2}{\sum_{i=1}^{m} (y_i - \bar{y})^2}
\]

\[
R^2 = 1 - \frac{\text{SS}_{\text{err}}}{\text{SS}_{\text{tot}}} = 1 - \frac{\sum_{i=1}^{m} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{m} (y_i - \bar{y})^2}.
\]

The third evaluation expressed the similarity and spatial distribution tendency of the overall pixel values between the prediction results output from the model, including the peak signal-to-noise ratio (PSNR) \([40]\) and structural similarity (SSIM) \([41]\)

\[
\text{PSNR} = 10 \times \log_{10} \left[ \frac{(2^n - 1)^2}{\text{MSE}} \right]
\]

\[
\text{SSIM} (x, y) = \frac{[2\mu_x\mu_y + (k_1L)^2][2\sigma_{xy} + (k_2L)^2]}{[\mu_x^2 + \mu_y^2 + (k_1L)^2][\sigma_x^2 + \sigma_y^2 + (k_2L)^2]}
\]

where \(\mu_x\) and \(\mu_y\) are the average of two images, \(\sigma_x\) and \(\sigma_y\) are the standard deviation of two images, \(\sigma_{xy}\) is the covariance of two images, \(L\) is the dynamic range of pixel values, and \(k_1\) and \(k_2\) are two scalar constants \((k_1 = 0.01\) and \(k_2 = 0.03\) by default).

### E. Model Training

For a neural network model with a recurrent structure, even if the parameter quantity is small, the resource consumption for computing and storage is relatively high during the training process. The weight item (i.e., convolution kernel parameters) in each channel of the ConvGRU layer is a 2-D matrix, and the feature map it handles has five dimensions. Thus, the overheads for all types of resources are much higher than those for traditional RNN networks. Moreover, the NDVI raster data used for model training and prediction in the experiment had a large size and a long time series. This further increased the requirements for workstation hardware. However, limited by cost and the experimental environment, we aimed for a balance in the experimental platform (see Table II). Table III gives the important software configurations for this article.

Across several experiments, considering the three aspects of model calculation efficiency, result accuracy, and hardware conditions in the final training parameters, the epoch was set at 128, the batch size was set at 1, the initial learning rate was set at 5 \(\times\) 10\(^{-4}\), and we selected adaptive motion (Adam) estimation \([42]\) as the optimizer. Considering the large feature map temporarily stored for the calculation process, we used a model with a half-precision (16-bit) floating point format for training. Moreover, we set a callback function to monitor when the loss of the validation set did not decrease for eight consecutive epochs and/or the learning rate decreased to 50% of the current value. The metric for the training process was set to SSIM and the custom loss function was the weighted sum of the logarithmic hyperbolic cosine function (LogCosh) and SSIM. The mathematical expression for the loss function is as follows:

\[
\text{Loss} = W_{\text{LogCosh}} \times \text{LogCosh}
\]

\[
+ W_{\text{SSIM}} \times (1 - \text{SSIM})
\]

where \(W_{\text{LogCosh}}\) and \(W_{\text{SSIM}}\) are the weights of LogCosh and SSIM, respectively, and \(W_{\text{LogCosh}} = 0.4\) and \(W_{\text{SSIM}} = 0.6\) are the settings for the experiment.

After 128 epochs on the training set, the network finally approximated convergence. We compared the training process for the four datasets mentioned above. The SSIM metrics for the network models reached 0.78–0.82 for the training set and 0.75–0.78 for the test set. In general, the performance of the unprocessed dataset was the worst, and the performance of the dataset after grouping tanh-ln fitting scaling was the best. The loss function values and metrics for the training process are shown in Fig. 8.

### IV. RESULTS AND DISCUSSION

#### A. Spatiotemporal Characteristics of the UHRB

To study the NDVI characteristics over time in the UHRB, we used the average values for the NDVI raster from 2000 to 2020 to represent vegetation coverage and growth in the region. We calculated the annual and monthly NDVI averages in the
TABLE III
IMPORTANT SOFTWARE (PACKAGE) CONFIGURATION

| Project   | Graphics Driver | CUDA   | CUDNN  | Python | Keras | TensorFlow |
|-----------|-----------------|--------|--------|--------|-------|-------------|
| Content   | 511.09 Studio   | V11.2.2| V8.1.1.33| 3.9.8  | 2.7.0 | 2.7.1       |

Fig. 8. Loss function and metrics performance in the training process: (a) Unprocessed dataset. (b) Scaling by fitted Fourier series. (c) Scaling by grouping the fitted univariate linear function. (d) Scaling by grouping the fitted tanh-ln function.

The changes in the annual average were linearly fitted, as shown in Fig. 9(a). The average annual NDVI value in the UHRB ranged from 0.19 to 0.23, with no obvious fluctuation. The overall growth trend was stable ($R^2 = 0.8174$), and the growth rate was 0.0017/a, indicating that the overall vegetation growth was good. Fig. 9(b) shows a box plot of the monthly NDVI average in the study area. The monthly average differences were obvious, and the changes revealed clear annual periodicity. The most significant differences were in June, July, and August (i.e., in the summer) in the study area.

We used the TS–MK method to produce an annual NDVI change map (see Fig. 10). The combined TS–MK test effectively reflected the spatial distribution characteristics of the NDVI change trend in the UHRB from 2000 to 2020. The proportions of vegetation in the UHRB showed increasing and decreasing trends of 93.7% and 6.3%, respectively. Most areas exhibited an increasing medium–low growth rate ($> 0/10^3 \cdot a$ and $< 6/10^3 \cdot a$), with only a few areas, mainly in the north and northeast corners of the UHRB, showing a high growth rate ($> 6/10^3 \cdot a$). About 42.94% of the areas, mainly in the southern fringe zone, did not pass the significance test ($p > 0.1$). Of all the areas, 34.52 showed extremely significant growth, mainly in the northern fringe zone and a few areas in the central region. Of all the areas, 13.83 showed a significant growth trend, with no concentration or clustering characteristics, but a wide distribution. Only a few areas showed significant and extremely significant declines scattered across the southern edge and eastern regions.

Research has shown that vegetation changes in the UHRB are mainly affected by climatic factors [14]. Changes in vegetation in the Heihe River Basin, according to the NDVI, are greatly affected by the combined influence of temperature and
precipitation [7]. In a relatively humid environment, the impact of temperature changes on the spatial distribution of vegetation may be amplified, and the feedback on the spatial distribution of vegetation may increase [12]. Also, the spatial distribution of precipitation largely determines the spatial distribution of water in the UHRB, and the growth of vegetation is affected by water conditions, with the vegetation characteristics reflecting the availability of water [14].

From 2000 to 2020, vegetation in the UHRB showed overall improvement, with only a few areas exhibiting degradation. This experimental result was consistent with a previous study on the spatial distribution and change trends in the NDVI in the UHRB [20], reflecting the reliability of the training datasets and the effective implementation of regional ecological protection policies.

B. Comparison of Prediction Results for Different Numerical Processing Methods

In this article, we used the original unprocessed NDVI raster data and the raster data processed using the three numerical processing methods discussed in Section III-B. for training and prediction, respectively. To select the most suitable numerical processing method for the dataset for model training and prediction, we needed to compare the prediction results and the labelled raster data of the test set in time steps corresponding to the NDVI raster data (23 time steps in the last year) using all the evaluation methods mentioned in Section III-D (see Table IV).

Moreover, we used K-fold cross-validation for the training of our model, and there were almost no significant differences in the evaluation metrics across multiple experiments. Therefore, the values for the evaluation metrics given in Table IV are the average values from multiple experiments using the test set. Table IV given that the results for the model training and prediction of the dataset after numerical processing by group tanh-ln fitting were better than the prediction results for the other three datasets using most evaluation metrics.

We also drew the metrics for the output results for each time step of the test set (see Fig. 11). According to the prediction results for the dataset after the numerical process using group tanh-ln fitting, the evaluation metrics for time steps in autumn and winter were significantly better than for the untreated data or the data processed using other methods.

Moreover, we calculated the NDVI annual average values, including the original time series NDVI raster data and the complete prediction results, using all the above-mentioned datasets, as shown in Fig. 11. In the subsequent prediction process, the prediction results from the dataset without numerical processing exhibited a gradual error accumulation due to the vanishing
TABLE IV

PERFORMANCE EVALUATION OF THE PREDICTION RESULTS USING DIFFERENT NUMERICAL PROCESSING METHODS

| Methods       | MAE    | MSE     | R²     | EVar   | SSIM ** | PSNR ** |
|---------------|--------|---------|--------|--------|---------|---------|
| Unprocessed   | 0.03931| 4.109×10⁻³| 0.6581 | 0.6972 | 0.7352  | 24.93   |
| Fourier Series| 0.03873| 4.076×10⁻³| 0.6903 | 0.6541 | 0.7404  | 24.96   |
| Linear (Grouped) | 0.03777| 3.892×10⁻³| 0.6866 | 0.7158 | 0.7335  | 24.49   |
| tanh-In (Grouped) | 0.03795| 3.877×10⁻³| 0.6916 | 0.7192 | 0.7434  | 25.13   |

*Note: The calculation results using SSIM in MATLAB, scikit-image, and TensorFlow may differ slightly. The calculation results from the TensorFlow deep-learning framework were used in this experiment.

**Note: Both SSIM and PSNR need to use the maximum dynamic pixel value range in the calculation process. The MODIS NDVI data range was [−0.2, 1], so the dynamic ranges were set to 1.2 for this experiment.

Fig. 11. Time series changes according to the evaluation metrics.
gradient, leading to an abnormal gradual decline in the annual average prediction. According to the prediction results from the datasets processed by Fourier series fitting and group univariate linear fitting, the predicted NDVI annual average showed an obvious upward acceleration. Therefore, the data after numerical processing using the group tanh-ln fitting method had better prediction performance than the unprocessed data or the data processed using other methods in terms of evaluation metrics and tendencies.

In summary, we selected the method of grouping the maximum TopK averages by scene number and tanh-ln respective curve fitting as the most suitable numerical processing method for predictions based on the study dataset.

C. Spatial and Temporal Distribution of Errors

Based on the best numerical processing method selected by comparing the above experimental results, we subtracted the output results for the test set from the corresponding labeled raster data and took the absolute value. We selected error diagrams for the 12 time steps from the test set and plotted them in Fig. 12. The spatial errors were generally evenly distributed, mostly as small values. The time steps with large absolute error values (red and yellow regions) existed in spring and autumn, when vegetation growth changed most obviously, and their spatial locations had little impact on the study area. In other words, after the numerical group tanh-ln fitting, the NDVI raster data were predicted using the neural network model based on the ConvGRU, and the results were reliable.

D. Spatiotemporal NDVI Variation Characteristics in the UHRB in the Future

We used the ConvGRU spatiotemporal prediction model to obtain NDVI images of the UHRB from 2020 to 2031, and then combined the TS and MK methods to analyse spatiotemporal trends in the vegetation. Regarding time, the NDVI for the UHRB from 2020 to 2031 showed an upward trend ($R^2 = 0.9729$) at a rate of 1.5$/10^3 \cdot \alpha$, indicating that the overall vegetation in the UHRB will improve in the future (see Fig. 13), consistent with the predictions of Zhang et al. [20] and Ding [43]. Based on the distribution of spatial changes, the increased area and the decreased area of vegetation in the UHRB accounted for 83.04% and 16.96%, respectively [see Fig. 14(a)]. This upward NDVI trend was regional, and the spatial change rate was mainly concentrated between 0 and $6 \times 10^{-3}$, accounting for 82.67% of the entire study area, and it was widely distributed across various areas of the UHRB. Of these areas, only 16.59% did not pass the significance test ($p > 0.1$), 60.95% showed a noteworthy upward trend [see Fig. 14(b)], and 8.17% showed a significant upward trend with a wider distribution. The spatial reduction NDVI trend was mainly concentrated between 0 and $-3 \times 10^{-3}$, distributed mostly around the Maying and Fengle Rivers, and these areas exhibited a noteworthy downward trend (about 8.04%).

The differences between these prediction results and those of Zhang et al. [20] and Ding [43] may be because Zhang et al. used the Hurst exponent to qualitatively predict future vegetation changes in the UHRB. This method can analyse increases or decreases in vegetation in the future but cannot predict the degree of vegetation change. Ding used a multiple linear regression model to predict future changes in vegetation in Northern China. Because the selected area was large and the selected factors were inadequate, the results were inconsistent and imprecise. The ConvGRU model with convolutional and cyclic structures used in this article accounts for both local and overall spatial characteristics, has high spatial accuracy, and
outputs the data from the prediction model continuously and intuitively. We accurately predict that the overall vegetation in the UHRB will significantly improve after 2020, but there will be a significant degradation trend around the Maying and Fengle Rivers in the basin. Therefore, the ConvGRU model is a suitable technology for predicting vegetation changes. It can accurately predict long-term vegetation changes and support corresponding active measures to ensure and improve vegetation conditions in the region.

V. CONCLUSION

In this article, we used MODIS NDVI data for the UHRB in the Qilian Mountains and constructed a neural network spatiotemporal prediction model based on a ConvGRU after scene-by-scene extreme-value tanh-log fitting and numerical scaling. We performed forecast and trend analyses on the original time series data and forecast results, respectively. The conclusions are as follows.

1) Compared with traditional methods based on sampling point regression predictions, the NDVI spatiotemporal prediction model we built has the advantages of complete direct grid prediction while ensuring smaller errors and higher accuracy, providing a new solution for such problems.
2) The spatiotemporal NDVI prediction method based on a ConvGRU can effectively capture spatiotemporal correlation characteristics and simultaneously display NDVI spatiotemporal changes in the study area completely and intuitively.
3) In the future, the overall vegetation in the UHRB will significantly improve, and only the Maying and Fengle River areas of the basin will show a significant degradation trend.

Our results demonstrate that neural networks can handle highly complex nonlinear models and can model a large number of different functions to solve complex spatial prediction problems. Under the combined influence of climate change and human activities, some vegetation areas in the UHRB in the Qilian Mountains may degrade in the future. Our proposed NDVI spatiotemporal prediction model, based on a ConvGRU, can effectively predict future changes in vegetation in the study area. This model can support large-scale NDVI prediction applications, considering both space and time, and can simultaneously ensure satisfactory accuracy. However, the modeling time for this model is relatively long, and the hardware requirements are relatively high. For a different study area, remodeling will be necessary. In future research, we will compare our model with other deep-learning models to identify the shortcomings of the proposed model as far as possible and adjust the model to improve the accuracy and efficiency of the prediction results.

ACKNOWLEDGMENT

The authors would like to thank all the funds and projects that have supported this work. The authors would like to thank other professors and schoolmates for their help in this work. The author would like to thank the two reviewers for their suggestions in this work, which make our research completer and more optimized. The authors would like to thank NASA Goddard Space Flight Center (GSFC) Level-1 and Atmosphere Archive & Distribution System (LAADS) Distributed Active Archive Center (DAAC) for providing access and download of MODIS NDVI data.

REFERENCES

[1] B. Li, W. Wang, L. Bai, N. Chen, and W. Wang, “Estimation of above-ground vegetation biomass based on landsat-8 OLI satellite images in the Guanzhong basin, China,” Int. J. Remote Sens., vol. 40, no. 10, pp. 3927–3947, Dec. 2018.
[2] S. Thomas, Climate Change 2013: The Physical Science Basis: Working Group I contribution to the Fifth assessment Report of the Intergovernmental Panel On Climate Change, Cambridge,U.K.: Cambridge Univ. Press, 2014.
[3] C. C. Barford et al., “Factors controlling long-and short-term sequestration of atmospheric CO2 in a mid-latitude forest,” Science, vol. 294, no. 5547, pp. 1688–1691, Nov. 2001.
[4] B. Zeng, “Vegetation responses to climate change on the tibetan plateau 1982-2003,” Ph.D. dissertation, Dept. Nat. Geo., Lanzhou Univ., Lanzhou, China, 2008.
[5] L. Kang, L. Di, M. Deng, Y. Shao, G. Yu, and R. Shrestha, “Use of geographically weighted regression model for exploring spatial patterns and local factors behind NDVI-precipitation correlation,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 7, no. 11, pp. 4530–4538, Nov. 2014.
[6] X. Yang, “Spatial-temporal variation of desert vegetation and its response to climate change in Hexi area during 1982-2013,” Ph.D. dissertation, Dept. Biogeogr., Lanzhou Univ., Lanzhou, China, 2015.
[7] L. Qiu, L. Zhang, Y. He, Z. Diao, and Y. Chen, “Remote sensing monitoring on vegetation dynamic change in Qilian mountain from 2000 to 2017,” Remote Sens. Inform., vol. 34, no. 4, pp. 97–107, Aug. 2019.
Yi He received the B.S. degree in geographic information system from Lanzhou Jiaotong University, Lanzhou, China, in 2011, and the Ph.D. degree in earth system science from Lanzhou University, Lanzhou, in 2016.

He has been a Postdoctoral Researcher with the School of Environment and Municipal Engineering, Lanzhou Jiaotong University. He is currently an Associate Professor with the faculty of geomatics, Lanzhou Jiaotong.

His research interests include disaster remote sensing, ecological remote sensing, image processing and time series InSAR prediction based on deep learning.

Sheng Yao received the B.E. degree in digital media technology from the North China University of Technology, Beijing, China, in 2020. He is currently working toward the M.S. degree with the Faculty of Geomatics, Lanzhou Jiaotong University, Lanzhou, China.

His research interests include the application of deep learning in remote sensing.

Shengpeng Cao received the B.E. degree in remote sensing science and technology in 2018 from the Lanzhou Jiaotong University, Lanzhou, China, where he is currently working toward the M.S. degree with the Faculty of Geomatics, Lanzhou Jiaotong University.

His research interest covers drought remote sensing monitoring and predicting.

Qiang Sun received the B.E. degree in surveying and mapping engineering from the Henan Polytechnic University, Jiaozuo, China, in 2018. He is currently working toward the M.S. degree with the Faculty of geomatics, Lanzhou Jiaotong University, Lanzhou, China.

His research interests include vegetative cover remote sensing monitoring and predicting.