Deep Learning for Daily 2-m Temperature Downscaling

Shuyan Ding[1], Xiefei Zhi[1], Yang Lyu[1], Yan Ji[1], and Weijun Guo[2]

1. Key Laboratory of Meteorology Disaster, Ministry of Education (KLME)/Joint International Research Laboratory of Climate and Environment Change (ILCEC)/Collaborative Innovation Center on Forecast and Evaluation of Meteorological Disasters (CIC-FEMD), Nanjing University of Information Science and Technology, Nanjing, Jiangsu 210044, China.

2. Xiamen Air Traffic Management Station of China Civil Aviation, Xiamen, Fujian 361006, China

Corresponding author: Xiefei Zhi (zhi@nuist.edu.cn)

Key Points:

- This paper presents a novel deep learning downscaling method, UNR-Net, capable of downscaling daily 2-m temperature by a factor of 10
- The overall performance of the UNR-Net method surpasses the U-Net method and linear regression method
- The 12 components-based error decomposition method is proposed to analyze the error source of different models.

Abstract

This study proposes a novel method, which is a U-shaped convolutional neural network that combines non-local attention mechanisms, Res2net residual modules, and terrain information (UNR-Net). The original U-Net method and the linear regression (LR) method are conducted as benchmarks. Generally, the UNR-Net has demonstrated promise in performing a 10x downscaling for daily 2-m temperature over North China with lead times of 1–7 days and shows superiority to the U-Net and LR methods. To be specific, U-Net and UNR-Net demonstrate higher Nash-Sutcliffe Efficiency coefficient（NSE) values compared to LR by 0.052 and 0.077, respectively. The corresponding improvements in pattern correlation coefficient are 0.013 and 0.016, while the root mean square error values are higher by 0.22 and 0.338, respectively. Additionally, the structural similarity index metric is higher by 0.033 and lower by 0.015. Furthermore, regions with significant errors are primarily distributed in complex terrain areas such as the Taihang Mountains, where UNR-Net exhibits noticeable improvements. In addition, the 12 components-based error decomposition method is proposed to analyze the error source of different models. Generally, the smallest errors are observed during the summer season and the sequence error component is proven to be the main source error of 2-m temperature forecasts. Furthermore, UNR-Net consistently demonstrates the lowest errors among all 12 error components. Therefore, combining the numerical weather prediction model and deep learning method is very promising in downscaling temperature forecasts and can be applied to routine forecasting of other atmospheric variables in the future.

Plain Language Summary

This research proposes a new method for downscaling using deep learning. The method uses a specific type of neural network called UNR-Net, which combines attention mechanisms, residual modules, and terrain information. The performance of UNR-Net is compared to two other methods: U-Net and LR. In the study, UNR-Net shows promise in performing a 10x downscaling of the daily 2-m temperature in North China. The UNR-Net demonstrates the best overall performance among all the comprehensive indicators (NSE, pattern correlation coefficient, root mean square error, and structural similarity index metric). Errors in the predictions are mainly found in complex terrain areas like the Taihang Mountains, but UNR-Net shows noticeable improvements in these regions. The study also proposes a 12 components-based error decomposition method to analyze the error sources of different models. All in all, it is found that

the smallest errors are observed during the summer season and the main source error is the sequence error component. Additionally, when considering lead times of 1–7 days, UNR-Net consistently shows the lowest errors among all 12 error components. Based on these findings, combining numerical weather prediction models with deep learning methods holds great promise for generating high-resolution temperature forecasts.

1. Introduction

Temperature is a meteorological element closely related to human life. With the advancement of society, there is an increasing demand for high-resolution temperature forecasts. However, in the present era, the resolution of numerical models is limited due to factors such as computational costs, scale sensitivity, and mismatches (Rind et al., 1992), which pose challenges in meeting the requirements of practical applications and scientific research. (Roberts et al., 2018; Feser et al., 2011; Wilby & Wigley, 1997). Therefore, downscaling methods have emerged.

These methods utilize appropriate refinement processes to infer meteorological element information at local scales based on the available low-resolution data (Höhlein et al., 2020). Due to the complexity of spatiotemporal characteristics, downscaling remains a challenging and intricate problem. Over the past few decades, various downscaling techniques have been proposed, including simple downscaling, dynamical downscaling (Jing et al. 2022; Wang et al. 2021), and statistical downscaling (Sharifi et al., 2019; Fowler et al., 2007). Among these, statistical downscaling exhibits a distinct advantage due to its high accuracy, excellent scalability, and lower computational resource requirements (Kim & Barros 2002; Frei et al., 2003; Hagemann et al., 2004; Ji et al., 2023a; Mannig et al., 2013).

In the past few decades, numerous advancements have been made in statistical downscaling techniques. Although traditional statistical approaches can to some extent enhance the resolution, they still have limitations in utilizing spatial and temporal dependencies, resulting in limited fitting capabilities (Chen et al., 2018; Wilby et al., 1998; He et al., 2016b). With the advent of the big data era, deep learning has the potential to discover features in high-dimensional data and capture the underlying nonlinear relationships between various meteorological variables (Yuan et al., 2020). It shows promise in terms of both accuracy and efficiency, surpassing previous methods (Höhlein et al., 2020). However, the use of deep learning methods in the field of meteorological downscaling is still in its early stages and faces challenges such as inadequate

description of complex features and poor performance in extreme events (Baño-Medina et al., 2020; Ji et al., 2022; Ji et al., 2023b; Vandal et al., 2019). Therefore, further practical exploration and research are needed to address these issues.

Presently, the field of deep learning offers numerous techniques that are well-suited for addressing challenges in the domain of downscaling. Due to its ability to incorporate receptive fields of varying sizes, U-Net has achieved success in semantic segmentation tasks (Ronneberger et al., 2015). Subsequently, it has also shown promising performance in tasks such as forecast calibration (Han et al., 2021; Zhu et al., 2022) and downscaling (Doury et al., 2023; Sha et al., 2020). However, when U-Net is employed for downscaling end-to-end tasks, the accuracy and practical effectiveness of the results can still be further improved through existing techniques. Mnih et al. (2014) achieved impressive results and gained widespread attention by incorporating attention mechanisms into convolutional neural networks for image processing tasks. Since then various attention mechanisms have emerged (Hu et al., 2018; Woo et al. 2018), and it has also found applications in downscaling (Park et al., 2022; Jing et al., 2022; Gerges et al., 2022). In theory, deeper networks have larger receptive fields, allowing them to integrate more information and potentially achieve better results. However, training deep networks can encounter challenges such as vanishing/exploding gradients and degradation (Pan et al., 2019). The Residual Network (ResNet), proposed by He et al. (2016a) successfully addressed the issue of network degradation. Subsequently, numerous studies have discussed the concept of residuals and proposed various variants (Xie et al., 2017; Huang et al., 2017; Veit et al., 2016). Furthermore, regarding the utilization of meteorological variables, apart from studies that solely utilize the meteorological variables at the downscaled scale (Kumar et al., 2021; Höhlein et al., 2020), some researchers have taken into account the physical significance and constraints of meteorological variables by incorporating terrain data into neural networks (Sha et al., 2020). But most studies that utilize multivariate data simply incorporate auxiliary data by stacking channels during input. This approach fails to effectively utilize higher-resolution auxiliary information compared to the target resolution, and there is also a lack of discussion regarding the optimal utilization of auxiliary information.

As a result, we propose a novel U-shaped convolutional neural network called UNR-Net, which integrates a nonlocal attention mechanism (Wang et al., 2018), Res2net (Gao et al.,2019), and terrain information for downscaling temperature. A non-local attention mechanism can allocate

importance to each position from a global perspective, considering the correlations between high-resolution observational data and low-resolution forecast data while disregarding distance. Res2net modules serve the purpose of not only mitigating model degradation issues but also efficiently coupling multiple receptive field sizes. We employ multiple convolutional operations to progressively decrease the resolution of the terrain data while increasing the number of channels and then fuse these products into the network through skip connections, which can not only adjust the feature map size of the high-resolution terrain data for easier input, but also preserve the high-resolution information of the terrain data, and control the proportion of terrain information in the network during feature extraction and downscaling. Additionally, in the downscaling part of the network, a combination of nearest-neighbor interpolation and convolution was utilized for upsampling (Dong et al., 2016), which can avoid the "checkerboard effect" caused by transpose convolution (Gauthier, 2014; Dumoulin et al., 2017).

The evaluation phase of the model after the modeling process is crucial. For model evaluation, the performance of UNR-Net is compared with LR, the original U-Net, and low-resolution forecast data with the comprehensive evaluation metrics NSE, pattern correlation coefficient (PCC), root mean square error (RMSE), and structural similarity index metric (SSIM). Still, the comprehensive evaluation metrics can only assess the overall performance of methods from various perspectives, which often lack detailed assessment and specific physical significance. Error decomposition, on the other hand, can provide a further evaluation of the results and enhance the interpretability of the methods (Hodson et al., 2021). Initially, the mean squared error (MSE) is decomposed into four seasons, and then the error for each season is further decomposed into three components, enabling a more detailed and specific analysis. The remainder of this paper follows the following structure. Section 2 describes the utilized data. Section 3 outlines the methods employed. Section 4 analyzes the performance of the three downsampling methods. Finally, a summary and discussion are presented in Section 5.

2. Data

The research area of this paper covers the geographical coordinates of (107° to 122.9°E, 31.1° to 47°N). The eastern part of the region is characterized by low-lying terrain and proximity to the ocean, while the western part features higher elevations and is situated inland. The region is predominantly mountainous with hilly terrain, exhibiting significant topographical variations.

Due to the significant impact of elevation on 2-m temperature and the region's importance as an agricultural area, downscaled modeling in such complex terrain is highly challenging yet holds substantial practical significance.

The forecast data used in this study is sourced from the Global Ensemble Forecasting System (GEFS) of the National Centers for Environmental Prediction (NCEP). The model resolution is 0.25° × 0.25°, and the data covers the geographical region of (105° to 124°E, 30° to 49°N). The dataset spans the period from 1 January 2010 to 31 December 2019, with daily initializations at 0000 UTC. The experiment incorporates 2-m temperature data with lead times of 1–7 days.

In this study, the observational data utilized the ERA5-Land dataset provided by the European Centre for Medium-Range Weather Forecasts (ECMWF). The dataset had a resolution of 0.1° × 0.1° and covered the time period from 2 January 2010 to 7 January 2020, with a focus on daily 2-m temperature data at 0000 UTC and covers the geographical region of (105° to 124.9°E, 29.1° to 49°N).

The topographic data utilized in this study was derived from the ETOPO1 dataset provided by the National Oceanic and Atmospheric Administration (NOAA). The dataset has a resolution of 1′ × 1′ and covers the geographical region of (105° to 124°59′E, 29°1′ to –49°N) (Figure 1).

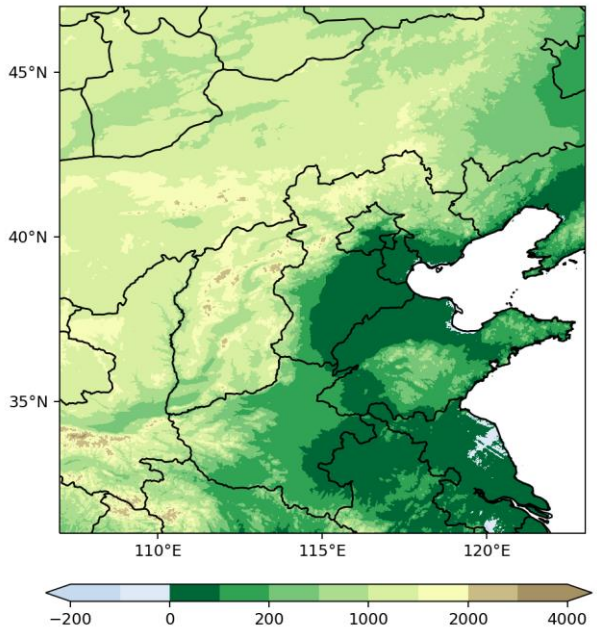


Figure 1. Study domain. The color bar represents the altitude of the terrain (m).

3. Methods

In this study, a total of three methods were employed to perform a 10x downscaling on the North China region. The three methods used were LR, U-Net, and UNR-Net. The first method, LR, is a traditional downscaling approach. It involves performing bilinear interpolation on the low-resolution forecast data and then applying linear regression. The second method is the unmodified original U-Net, where the upsampling process utilizes transpose convolutions. The third method is a modified version of the U-Net that combines non-local attention with Res2net residual modules. Additionally, it incorporates terrain information to enhance the performance of the model. These three methods utilize the data that has undergone data preprocessing as described in Section 3.1, and these three methods share the same training set, validation set, and test set. The two deep learning methods both employ an end-to-end approach, where the networks simultaneously perform the calibration and downscaling tasks. Additionally, the training process for both networks is identical and follows a supervised learning approach. The Adam optimizer is used for training in both cases (Kingma & Ba, 2017). The loss function used is MSE, defined as follows:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - y_i'\right)^2, \quad (1)$$

where N represents the number of grid points in a batch, i represents the grid point position, y_i represents the ground truth values, and y_i', represents the predicted values. The learning rate is set to 0.001, and it is decayed every 20 steps with a decay rate of 0.5. To avoid overfitting, early stopping is implemented to determine the stopping epoch. Finally, an evaluation is conducted on the results of the three methods, which includes the error between the high-resolution downscaled results and high-resolution observation, as well as the error between the low-resolution observation and the low-resolution forecast data after second-order conservative remapping scheme (Jones, 1999). Subsequently, an error decomposition is performed to further analyze the performance of the three methods.

3.1 Data Preprocessing

First, data processing is conducted on the forecast data. The forecast data has a resolution of 0.25° × 0.25°, and due to the downscaling factor of 10 used in the experiment, the resolution of the observational data is 0.1° × 0.1°. Therefore, the first step is to perform a second-order

conservative remapping scheme (Jones, 1999) to adjust the resolution of the forecast data to 1° × 1°. Due to the absence of observational data for the oceanic region, the oceanic part of the forecast data is assigned empty values. Then, the land data is standardized using the following formula:

$$x_{new} = \frac{x - \mu}{\sigma}, \tag{2}$$

where x represents the previous value of the data, x_{new} represents the new value of the data, μ represents the mean of the respective matrix, and σ represents the variance. Afterward, the oceanic region is filled using the nearest-neighbor interpolation method, as described by the formula:

$$f(i) = f(i_{nearest}), \tag{3}$$

where i represents the grid point location, $i_{nearest}$ represents the closest grid point location to i, $f(i_{nearest})$ represents the value of the nearest grid point, and $f(i)$ represents the value of the grid point i. By applying the mentioned processing to the forecast data, the low-resolution forecast data is obtained.

Next, the observational data are processed in a similar manner as the forecast data, including standardization and nearest-neighbor interpolation for the oceanic regions. This results in obtaining the high-resolution observational data.

Both the low-resolution forecast data and the high-resolution observational data are divided into training sets, validation sets, and testing sets. The training set consists of data with forecast start dates from 1 January 2010 to 31 December 2017. The validation set comprises data with forecast start dates from 1 January 2017 to 31 December 2018. Lastly, the testing set includes data with forecast start dates from 1 January 2018 to 31 December 2019.

Finally, the terrain data is processed using the same methods as applied to the forecast data, including standardization and nearest-neighbor interpolation for the oceanic regions. This results in obtaining the high-resolution terrain data.

3.2 Downscaling methods

3.2.1 LR

First, the low-resolution forecast data is subjected to bilinear interpolation. Bilinear interpolation involves performing linear interpolation in two directions. The formulas for bilinear interpolation are as follows:

$$f(X_r, Y_a) \approx \frac{X_b - X_r}{X_b - X_a} f(X_a, Y_a) + \frac{X_r - X_a}{X_b - X_a} f(X_b, Y_a), \tag{4}$$

$$f(X_r, Y_b) \approx \frac{X_b - X_r}{X_b - X_a} f(X_a, Y_b) + \frac{X_r - X_a}{X_b - X_a} f(X_b, Y_b), \tag{5}$$

where $f(X_a, Y_a)$, $f(X_b, Y_a)$, $f(X_a, Y_b)$, and $f(X_b, Y_b)$ represent the values of the four points in the respective directions. X_a, X_b, Y_a, and Y_b represent the positions on the coordinate axes, while $f(X_r, Y_a)$ and $f(X_r, Y_b)$ represent the points obtained through linear interpolation in the x-direction. After obtaining $f(X_r, Y_a)$ and $f(X_r, Y_b)$, linear interpolation is performed in the y-direction to obtain the value of the unknown point as follows:

$$f(X_r, Y_r) \approx \frac{Y_b - Y_r}{Y_b - Y_a} f(X_r, Y_a) + \frac{Y_r - X_a}{Y_b - Y_a} f(X_r, Y_b), \tag{6}$$

where $f(X_r, Y_r)$ represents the value of the unknown point. As a result, we obtain the forecast data with a resolution of 0.1° × 0.1°. Next, linear regression is applied to the forecast data, following the formula:

$$y_t = a + bx_t, \tag{7}$$

where y_t represents the predicted result, x_t represents the values for linear regression, and a and b are the coefficients of the linear regression. The coefficients a and b in linear regression are calculated using the following formulas:

$$a = \bar{y} - b\bar{x}, \tag{8}$$

$$b = \frac{\Sigma_{t=1}^{n} x_t y_t - n\bar{x}\bar{y}}{\Sigma_{t=1}^{n} x_t^2 - n\bar{x}^2}, \tag{9}$$

where x_t represents the forecast data in the training set, y_t represents the observational data in the training set, $\bar{x}$ represents the mean of the forecast data in the training set, and $\bar{y}$ represents the mean of the observational data in the training set. After obtaining the regression coefficients a

and *b* for each grid point, the regression equation is applied to the testing set. This process yields the downscaled results using the LR method for the testing set.

3.2.2 U-Net

The U-Net is a convolutional neural network architecture with a U-shaped structure that was first introduced in the field of semantic segmentation (Ronneberger et al. 2015). Its encoder-decoder structure is well-suited for handling pixel-level problems and is also well-suited for processing grid-based meteorological data. The downscaling U-Net architecture employed in this study employs multiple upsampling operations in the decoder section, iteratively generating feature maps at the desired target resolution (Figure 2).

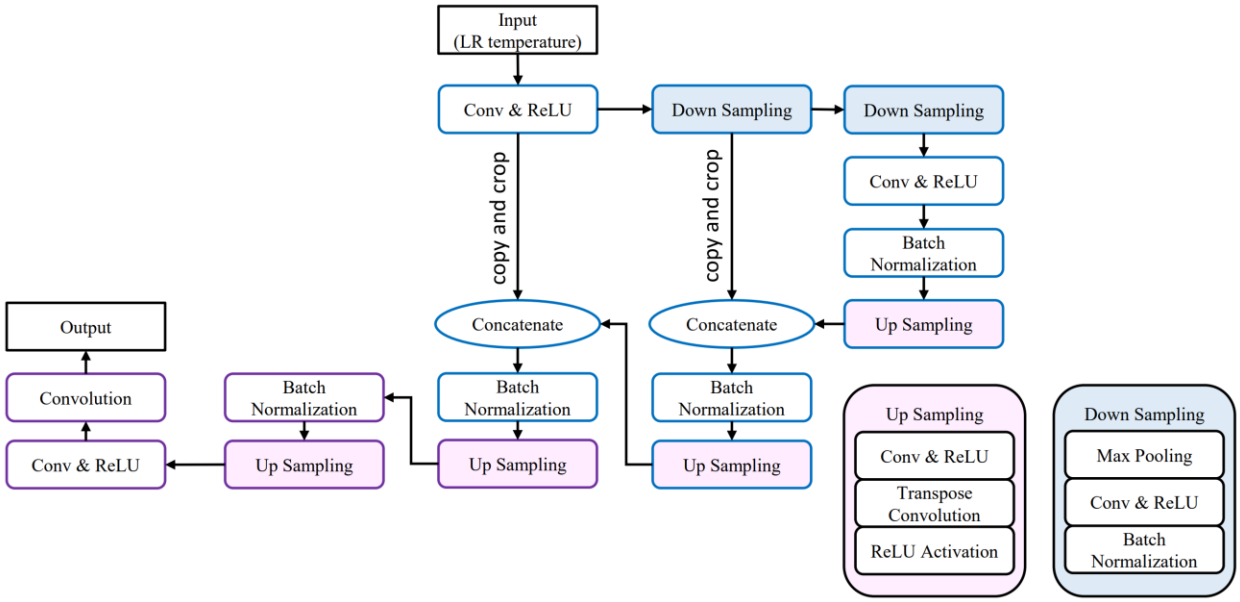


Figure 2. U-Net architecture. The rectangular box with right angles represents the data, while the rounded rectangles and ellipses represent different operations or functions. The text inside the rounded rectangles and ellipses indicates the name of the respective operation or function. The solid arrows indicate the flow of data. The blue box represents the feature extraction section, while the purple box represents the downscaling section.

The downscaling U-Net architecture in this paper consists of two components: the blue box representing the feature extraction section and the purple box representing the downscaling section. Placing the upsampling process after the feature extraction section can reduce computational load without compromising the accuracy of the results. The feature extraction

section accepts input data of low resolution for forecasting. The structure of the feature extraction section consists of the left encoder part and the right decoder part. Downsampling is a critical process in the encoder, and the downsampling module comprises pooling, convolution, and activation operations. Through the downsampling module, the size of the feature maps is halved, while the number of channels is doubled compared to the original. Upsampling is an essential process in the decoder, and the upsampling module consists of convolution, activation, transposed convolution, and activation operations. Through the upsampling module, the size of the feature maps is doubled, while the number of channels is halved compared to the original. Then the feature maps of the same size are concatenated through skip connections. Skip connections allow for the integration of information from different scales, thereby enhancing the network's ability to capture complex patterns and improve its fitting capability. The number of convolutional kernels in each layer of the feature extraction section is {16, 32, 64, 64, 64, 32, 32, 16}. The downsampling section includes two upsampling processes. The order of operations in the upsampling module is consistent with the feature extraction section. However, in the final upsampling module, the size of the feature maps is increased by a factor of 5. After the upsampling, batch normalization, convolution, and activation operations in the downsampling section, the high-resolution downscaled results are obtained. The number of convolutional kernels in each layer of the feature extraction section is {16, 16, 16, 16, 16, 16, 1}. The relationship to control the output feature map size in convolutional operations within the network can be represented as,

$$o = \left\lfloor \frac{i + 2p - k}{s} \right\rfloor + 1. \tag{10}$$

The relationship to control the output feature map size in transpose convolutional operations can be represented as:

$$o = s(i - 1) + k - 2p, \tag{11}$$

where o represents the size of the output feature map, i represents the size of the input feature map, p represents the padding size, k represents the size of the convolutional kernel, and s represents the stride of the convolution operation. The activation function used is Rectified Linear Unit (ReLU) (Glorot et al., 2011), which is widely used and highly effective in regression problems. Its formula is as follows:

$$ReLU(x_i) = max(0, x_i), \tag{12}$$

where, x_i represents the elements of the feature map to the activation function. Batch normalization is also employed in the network. Batch normalization helps alleviate the problem of internal covariate shifts during the training process (Ioffe & Szegedy, 2015). It stabilizes the learning process by normalizing the inputs of each layer within a mini-batch and has shown good performance in accelerating the training of various deep learning models and regularization (Silver et al., 2017). The formula for batch normalization is as follows;

$$BN(X) = \frac{X - E[X]}{\sqrt{\mathrm{Var}[X] + \epsilon}} \times \gamma + \beta, \tag{13}$$

where γ and β are trainable parameters, ϵ is a small constant value, X represents the feature map matrix, $E[X]$ is the mean of the feature map matrix X, and $\mathrm{Var}[X]$ is the variance of the feature map matrix X.

3.2.3 UNR-Net

The UNR-Net consists of three components: the auxiliary information processing section, the feature extraction section, and the downscaling section (Figure 3). The auxiliary information processing section receives high-resolution terrain data and outputs to both the feature extraction section and the downscaling section. The feature extraction section takes inputs from the low-resolution forecast data and the auxiliary information processing section, and outputs to the downscaling section. The downscaling section receives inputs from both the auxiliary information processing section and the feature extraction section, ultimately generating high-resolution downscaled results.

The auxiliary information processing section accepts high-resolution auxiliary data, and the resolution of the auxiliary data can be higher than the target resolution of the downscaling task. Because in this section, we control the size of the output feature maps by utilizing (Equation 9), the result is a gradual reduction in the dimensions of the feature maps. In this task, the resolution of the terrain data is 60 times higher than that of the low-resolution forecast data. In this network, the size of the feature maps gradually decreases by factors of 2, 3, 5, and 2. Specifically, the data that undergoes a total reduction of 6 and 30 times is input into the downscaling part, while the data that undergoes a total reduction of 60 times is input into the feature extraction part. The number

of convolutional kernels in this part is {8, 8, 8, 8, 6}, and after the first and second copies, the number of convolutional kernels for the subsequent convolution operations is {1, 1}.

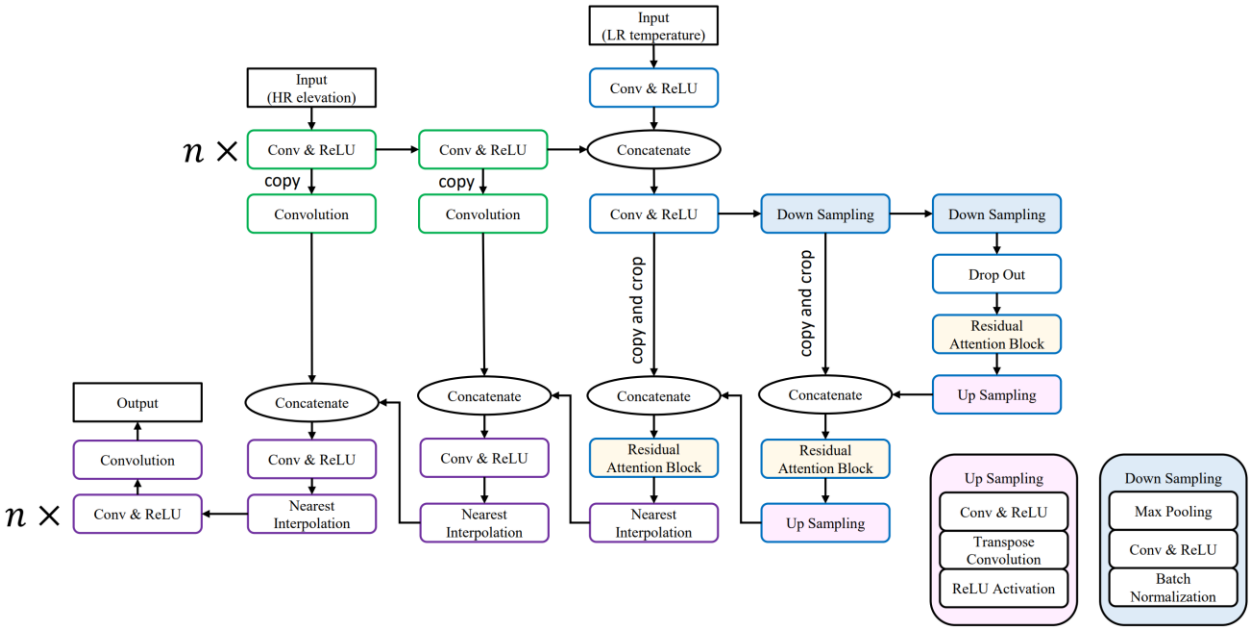


Figure 3. UNR-Net architecture. The green box represents the auxiliary information processing section, the blue box represents the feature extraction section, and the purple box represents the downscaling section. The rest of the instructions are the same as in Figure 2.

The difference between the feature extraction part in this section and the feature extraction part in the downscaling U-Net described in Section 3.2.2, is the utilization of dropout layers to prevent alleviate overfitting (Srivastava et al., 2014), the removal of some batch normalization operations (Li et al. 2019), and the addition of residual attention modules. The addition of the residual attention modules is one of the key improvements in the network.

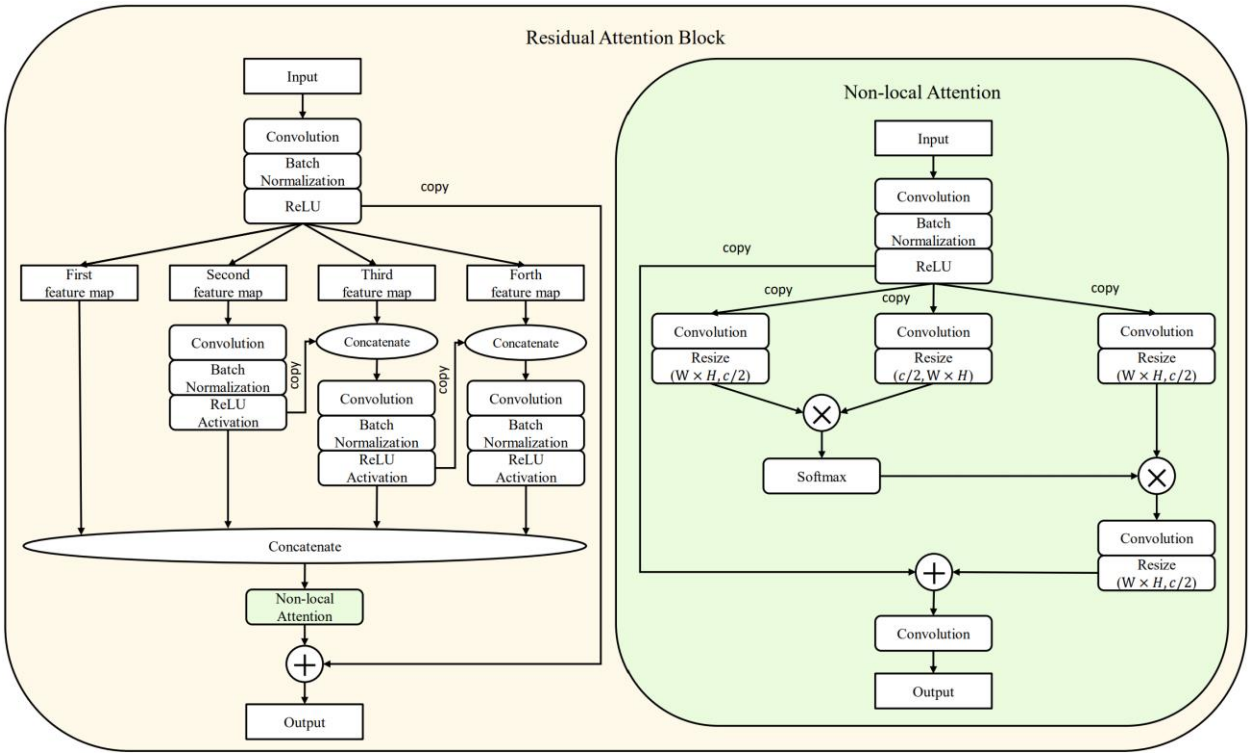


Figure 4. The Residual Attention Module. The numbers within the parentheses for the resize operation represent the shape of the output feature map, and the ⊗ represents matrix multiplication.

This module primarily consists of the Res2net module (Gao et al., 2019) and a non-local attention mechanism (Wang et al., 2018) (Figure 4). The Res2net module divides the feature map into four segments along the channel dimension and processes them separately. This approach enables more efficient integration of multiple receptive field sizes, allowing the model to capture information from different scales effectively. Additionally, it helps prevent model degradation, ensuring stable and robust performance. The nonlocal attention mechanism is a spatially sensitive attention mechanism. Within the attention part of the module, the input is first replicated four times for separate operations. One of the replicas is used for residual connection, while the other two replicas are used to generate attention weights. These attention weights are then multiplied with the fourth replica, resulting in a feature map that emphasizes important regions. During the generation of attention weights, the module performs matrix multiplication on the resized feature maps to produce the attention weights, similar to the process of generating a covariance matrix. These attention weights are then activated using the softmax activation function, given by the formula:

$$Softmax(z_i) = \frac{e^{z_i}}{\sum_{c=1}^{C} e^{z_c}}, \tag{14}$$

where z_i represents the elements of the attention weight matrix, and C represents the total number of elements in the attention weight matrix. Finally, the attention-weighted feature map is obtained by multiplying it element-wise with the fourth copy of the feature map that underwent convolution and resizing operations. This results in a feature map where the importance is allocated based on the attention weights. Therefore, this non-local attention mechanism can allocate the importance of each position in the feature map from a global perspective, considering the correlation between high-resolution observational data and low-resolution forecast data, while ignoring the spatial distance and capturing the interactions across different locations. The distinction between the downscaling components of UNR-Net and U-Net lies in the source of data and the methodology employed for upsampling. The downscaling section of this network not only receives inputs from the feature extraction part but also incorporates inputs from the auxiliary information processing section after each upsampling step. These input feature maps have a smaller receptive field, allowing the network to incorporate information from a smaller scale and enhance its fitting capability. Moreover, the upsampling method in the downscaling section of UNR-Net has been changed from transpose convolution to a combination of nearest-neighbor interpolation and convolution. This change was made to address the issue of checkerboard artifacts that can be introduced by transpose convolution, which can potentially affect the practical utility of the network (Dumoulin et al., 2017). The combination of nearest-neighbor interpolation and convolution for upsampling helps to avoid the occurrence of checkerboard artifacts without compromising result accuracy. The formula for nearest-neighbor interpolation is as follows:

$$srcX = dstX \times \frac{srcWidth}{dstWidth}, \tag{15}$$

$$srcY = dstY \times \frac{srcHeight}{dstHeight}, \tag{16}$$

where $dstX$ and $dstY$ represent the coordinates of the enlarged feature map's grid points; $dstWidth$ and $dstHeight$ represent the length and width of the enlarged feature map; $srcX$ and $srcY$ represent the coordinates of the original feature map's grid points; $srcWidth$ and $srcHeight$ represent the length and width of the original feature map.

The network shares the same operations and parameters as the U-Net, except for the differences mentioned above.

3.3 Evaluation metrics

We have employed six evaluation methods to assess the strengths and weaknesses of the three approaches from various perspectives. The six evaluation methods are the NSE (Nash et al., 1970), PCC, RMSE, SSIM, Root Mean Square Error Skill Score (RMSESS), and MSE. Their formulas are given as follows:

$$NSE = 1 - \frac{\Sigma_{i=1}^{n}(o_i - f_i)^2}{\Sigma_{i=1}^{n}(o_i - \bar{o})^2}, \tag{17}$$

$$PCC = \frac{\sum_{i=1}^{m}(f_i - \bar{f})(o_i - \bar{o})}{\sqrt{\sum_{i=1}^{m}(o_i - \bar{o})^2 (f_i - \bar{f})^2 \sum_{i=1}^{m}(o_i - \bar{o})^2}}, \tag{18}$$

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(f_i - o_i)^2}, \tag{19}$$

$$SSIM = \frac{(2\bar{o}\bar{f} + C_1)(2\sigma_{of} + C_2)}{(\bar{f}^2 + \bar{o}^2 + C_1)(\sigma_f^2 + \sigma_o^2 + C_2)}, \tag{20}$$

$$RMSESS = \frac{RMSE_{ref} - RMSE}{RMSE_{ref}}, \tag{21}$$

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(f_i - o_i)^2, \tag{22}$$

where o represents the observation; f represents the forecast; n represents the number of time steps; m represents the number of grid points; $\bar{o}$ and $\bar{f}$ represent the mean of observations and forecasts (in the case of NSE, it represents the mean over all time steps for a grid point, and in other formulas, it represents the mean over all grid points for a time step); respectively; σ_o is the standard deviation of observations; σ_f is the standard deviation of forecasts; σ_{of} is the covariance between observations and forecasts; C_1 and C_2 are constants. Among the six evaluation metrics, higher values indicate better performance, except for RMSE and MSE, in

which a lower value is desirable. NSE reflects the fitting effect of the downscaled results to the observations (Gerges et al., 2022). PCC measures the correlation between the forecast field and the observation field. $RMSE$ represents the absolute error between the forecast values and the observed values. $SSIM$ measures the similarity between the forecast field and the observation field, which is more in line with human visual perception. $RMSESS$ quantifies the comparison between two methods. MSE, as an absolute error, can amplify the $RMSE$.

3.4 Error decomposition

While comprehensive evaluation metrics allow for the quantification of model performance from various perspectives, they often lack interpretability and provide limited guidance for model improvement (Zhu et al., 2022). Error decomposition, on the other hand, enables the breakdown of absolute errors into interpretable components, facilitating a more in-depth evaluation (Hodson et al., 2021).

First, the model error $\boldsymbol{\epsilon}$ is defined as,

$$\boldsymbol{\epsilon} = \boldsymbol{f} - \boldsymbol{o}, \tag{23}$$

where $\boldsymbol{f}$ represents a downscaled results vector with n values, and $\boldsymbol{o}$ represents an observation vector with n values.

The error can be divided into four components representing the four seasons,

$$\boldsymbol{\epsilon} = \boldsymbol{\epsilon_1} + \boldsymbol{\epsilon_2} + \boldsymbol{\epsilon_3} + \boldsymbol{\epsilon_4}, \tag{24}$$

$$= \boldsymbol{\delta_1}\boldsymbol{\epsilon} + \boldsymbol{\delta_2}\boldsymbol{\epsilon} + \boldsymbol{\delta_3}\boldsymbol{\epsilon} + \boldsymbol{\delta_4}\boldsymbol{\epsilon}, \tag{25}$$

where 1, 2, 3, and 4 represent the four seasons (spring, summer, autumn, and winter, respectively), and $\boldsymbol{\delta}$ is a matrix of the same shape as $\boldsymbol{\epsilon}$, consisting of elements 0 and 1,

$$\delta_{ij} = \begin{cases} 1 \; if \; \epsilon_i \in season \; j \\ 0 \; otherwise \end{cases}. \tag{26}$$

So, the MSE can be decomposed as follows:

$$MSE(\boldsymbol{\epsilon}) = MSE(\boldsymbol{\epsilon_1} + \boldsymbol{\epsilon_2} + \boldsymbol{\epsilon_3} + \boldsymbol{\epsilon_4}), \tag{27}$$

$$= \frac{1}{n}\sum_{i=1}^{n}(\epsilon_{1i} + \epsilon_{2i} + \epsilon_{3i} + \epsilon_{4i})^2, \tag{28}$$

$$= \frac{1}{n}\sum_{i=1}^{n}(\epsilon_{1i})^2 + \frac{1}{n}\sum_{i=1}^{n}(\epsilon_{2i})^2 + \frac{1}{n}\sum_{i=1}^{n}(\epsilon_{3i})^2 + \frac{1}{n}\sum_{i=1}^{n}(\epsilon_{4i})^2$$
$$+\frac{2}{n}\sum_{i=1}^{n}(\epsilon_{1i}\cdot\epsilon_{2i}) + \frac{2}{n}\sum_{i=1}^{n}(\epsilon_{3i}\cdot\epsilon_{4i}) + \frac{2}{n}\sum_{i=1}^{n}(\epsilon_{1i}+\epsilon_{2i})\cdot(\epsilon_{3i}+\epsilon_{4i})\,. \quad (29)$$

Since the errors for the four seasons are orthogonal to each other, the term $\frac{2}{n}\sum_{i=1}^{n}(\epsilon_{1i}\cdot\epsilon_{2i}) + \frac{2}{n}\sum_{i=1}^{n}(\epsilon_{3i}\cdot\epsilon_{4i}) + \frac{2}{n}\sum_{i=1}^{n}(\epsilon_{1i}+\epsilon_{2i})\cdot(\epsilon_{3i}+\epsilon_{4i})$ equals zero. According to Hodson et al. (2021), each MSE can be decomposed into three components: Bias, Sequence, and Distribution. Therefore, it follows,

$$\begin{aligned} MSE(\boldsymbol{\epsilon}) = {} & Bias(\boldsymbol{\epsilon_1})^2 + Sequence(\boldsymbol{\epsilon_1}) + Distribution(\boldsymbol{\epsilon_1}) \\ & + Bias(\boldsymbol{\epsilon_2})^2 + Sequence(\boldsymbol{\epsilon_2}) + Distribution(\boldsymbol{\epsilon_2}) \\ & + Bias(\boldsymbol{\epsilon_3})^2 + Sequence(\boldsymbol{\epsilon_3}) + Distribution(\boldsymbol{\epsilon_3}) \\ & + Bias(\boldsymbol{\epsilon_4})^2 + Sequence(\boldsymbol{\epsilon_4}) + Distribution(\boldsymbol{\epsilon_4}). \end{aligned} \quad (31)$$

The Bias component quantifies the model's ability to accurately replicate the mean of the observations. The Sequence component measures the model's ability to accurately reproduce temporal sequences of events. The Distribution component quantifies the model's ability to accurately replicate the distribution of the observations. As a result, we have obtained the differences in terms of bias, sequence, and distribution across various seasons.

4. Results

4.1 General downscaling performance

Figure 5 illustrates the error performance of the downscaled results from the three methods and the low-resolution forecast data compared to the observations. the graph illustrates that the downscaled results from the three methods exhibit significantly better performance in terms of error compared to the low-resolution forecast data.

For NSE, PCC, and RMSE, the average differences between the LR method and U-Net over a lead time from 1 to 7 days are 0.052, 0.013, and 0.220, respectively. In comparison, the corresponding differences between the LR method and UNR-Net are 0.077, 0.016, and 0.338, respectively. The LR method exhibits the poorest performance. As for the two deep learning methods, their average differences are 0.025, 0.005, and 0.117, indicating a similar performance. Moreover, as the lead time increases, the differences among the three methods become more

pronounced. When the lead time is 1 day, the gap between U-Net results and low-resolution data is 114% (NSE), 111% (PCC), and 125% (RMSE) compared to the gap between LR results and low-resolution data. Similarly, the gap between UNR-Net results and LR results is 117% (NSE), 112% (PCC), and 131% (RMSE) compared to the gap with low-resolution data. When the lead time is 7 days, the gap between U-Net results and low-resolution data is 129% (NSE), 119% (PCC), and 141% (RMSE) compared to the gap between LR results and low-resolution data. Likewise, the gap between UNR-Net results and LR results is 142% (NSE), 127% (PCC), and 165% (RMSE) compared to the gap with low-resolution data. UNR-Net demonstrates a greater advantage as the lead time becomes longer.

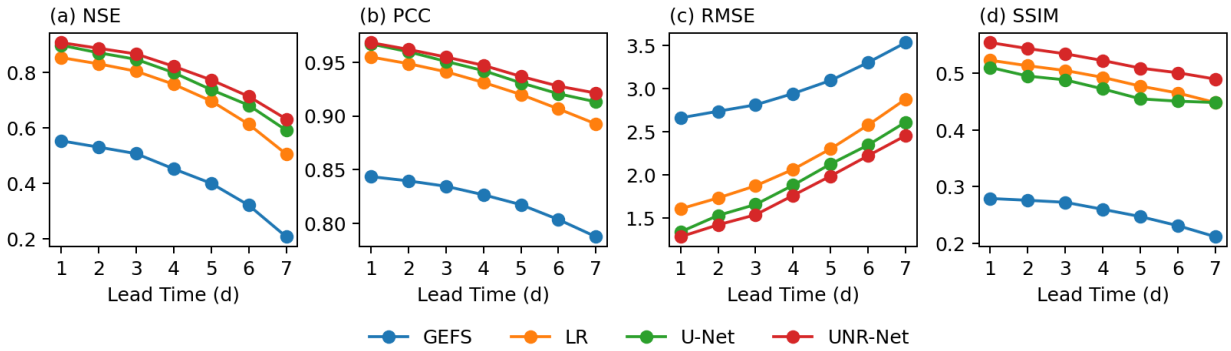


Figure 5. The overall evaluation of the three methods and the GEFS data. Variations in (a) NSE, (b) PCC, (c) RMSE, and (d) SSIM of 2-m temperature at lead times of 1–7 days derived from the GEFS, LR, U-Net, and UNR-Net averaged over North China.

However, in the case of SSIM, the average score of U-Net is even lower than that of the LR method by 0.015. This indicates that U-Net performs worse than the LR method when considering SSIM. Indeed, the other three evaluation metrics (NSE, PCC, and RMSE) unequivocally attest to the superior calibration ability of U-Net compared to the LR method. However, the evaluation of SSIM measures the degree of structural similarity, which suggests that U-Net does not excel in the downscaling task. On the other hand, the modified network UNR-Net, when compared to the LR method, exhibits significant improvement. This implies that the checkerboard artifacts introduced by the transposed convolutions used in the upsampling process of U-Net diminish the practical application value of its results. Conversely, the modified UNR-Net incorporates a combination of nearest-neighbor interpolation and convolution. This

approach successfully mitigates the issue of checkerboard artifacts, all the while maintaining unblemished result accuracy.

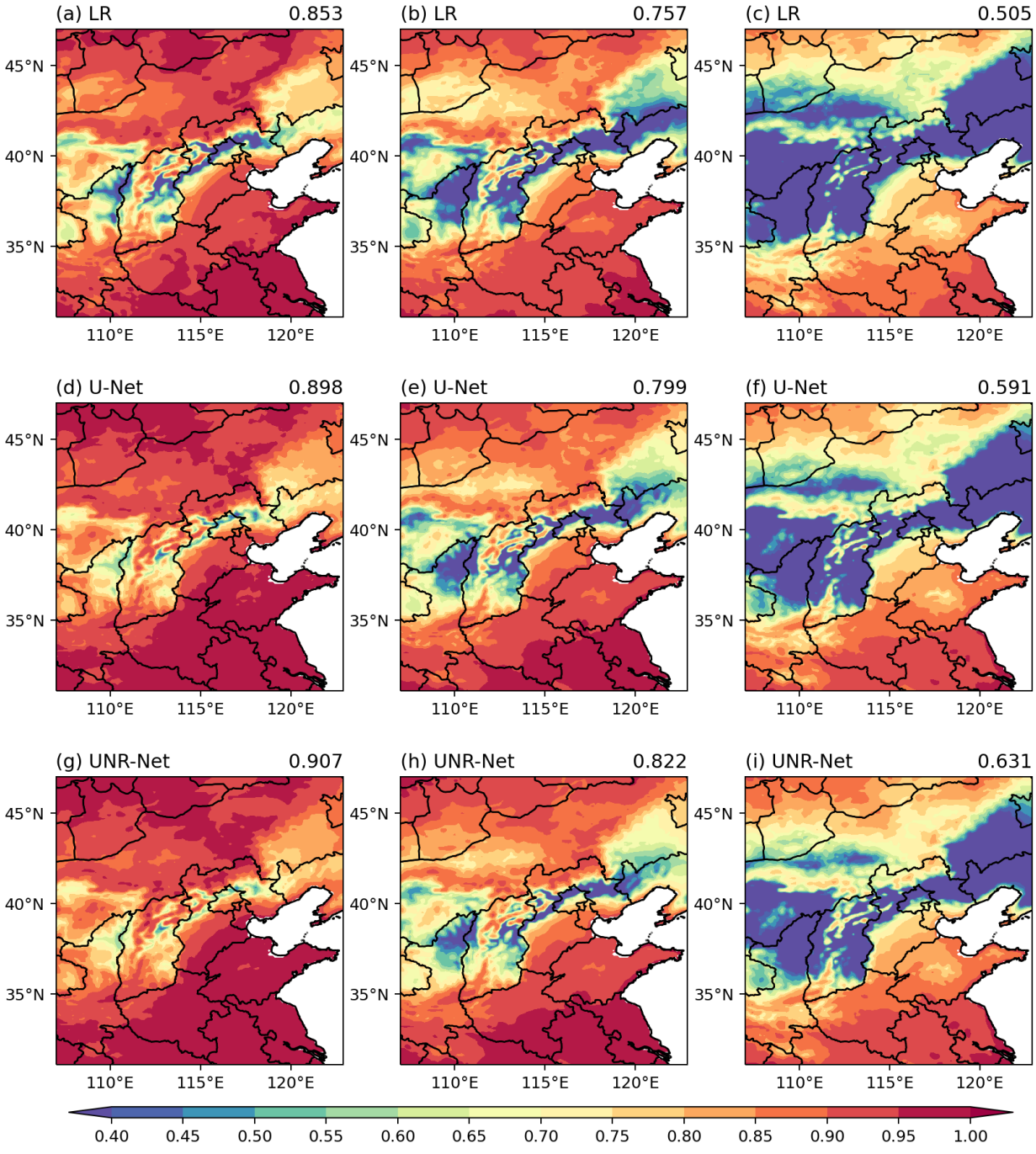


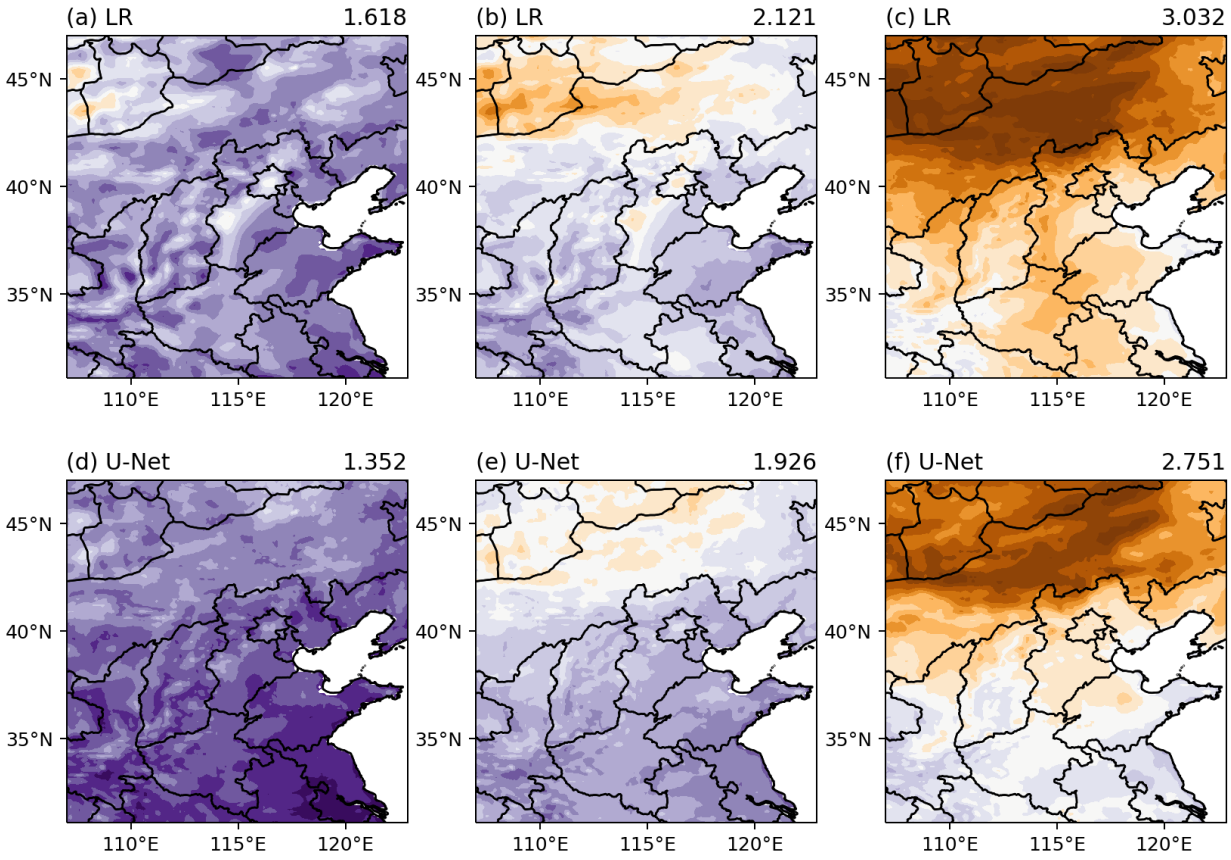
Figure 6. Spatial distributions of the NSE for 2-m temperature with lead times of 1, 4, and 7 days derived from LR (a–c), U-Net (d–f), and UNR-net (g–i), the values in the upper-right title represent the mean of NSE in each case.

The spatial distribution of NSE for the three methods is illustrated in Figure 6. Upon examining the spatial distribution of the downscaled results from the three methods in fitting high-resolution observational data, it becomes evident that the underperforming areas are concentrated around the Taihang Mountain range. In certain regions, the scores even dip below 0.45. The scores exhibit an increasing trend from both sides of the Taihang Mountain range in terms of spatial distribution. Overall, there is a pattern of higher scores in the southeastern and northwestern regions.

From LR to U-Net and then to UNR-Net, the fit of the predicted field to the observed field in the region of the Taihang Mountains shows improvement. Moreover, within all lead times depicted in the figure, the area characterized by scores below 0.45 exhibits a progressively diminishing extent, with notable improvements observed, particularly in the vicinity of the Taihang Mountains, for the two employed deep learning methodologies. Additionally, it is noteworthy that the area encompassing scores exceeding 0.95 demonstrates a gradual increase in size. When considering a lead time of 1 day, the distribution of regions with NSE scores below 0.45 for the LR method is concentrated in the northeastern part of Shaanxi province and in the vicinity of the Taihang Mountains. In fact, a majority of these areas even exhibit scores below 0.4. On the other hand, regions with NSE scores surpassing 0.95 are mainly limited to Jiangsu, Anhui, Hubei, and other areas. For the U-Net method, the extent of regions with NSE scores below 0.45 has been significantly reduced. In the northeastern part of Shaanxi province, there are no longer any areas with scores below 0.4. Furthermore, regions with NSE scores exceeding 0.95 now include Shandong and Henan. As for the UNR-Net method, the area with NSE scores below 0.45 has decreased compared to U-Net. In the northeastern part of Shaanxi province, southern Shanxi, and Liaoning province, there are no longer any regions with scores below 0.45. The regions near the Taihang Mountains with scores below 0.45 appear sporadically. Additionally, the region in the southeast of Hebei province is now encompassed within the area with NSE scores surpassing 0.95, and the overall area in Shandong with scores below 0.95 has decreased compared to U-Net. When considering lead times of 4 and 7 days, although the overall scores of all three methods have decreased, certain patterns still emerge. The LR method exhibits the largest area with scores below 0.45 and the smallest area with scores above 0.95. Conversely, UNR-Net showcases the smallest area with scores below 0.45 and the largest area with scores above 0.95. Especially noteworthy is the 7-day lead time, where the LR method has virtually no areas with scores above

0.95. In contrast, both deep learning methods still exhibit some distribution in the southeastern region. Furthermore, the area with scores above 0.8 for UNR-Net is notably larger than that for U-Net.

As the lead time increases, NSE scores progressively decrease. For instance, when comparing the 7-day lead time to the 1-day lead time, the LR method experiences a decrease of 0.348 in NSE, while the U-Net method decreases by 0.307 and the UNR-Net method decreases by 0.276. It is evident that UNR-Net exhibits a lower reduction magnitude compared to the other two methods. Additionally, UNR-Net maintains a larger area with high NSE scores. Therefore, UNR-Net demonstrates its superiority particularly in longer lead times, showcasing its favorable performance.



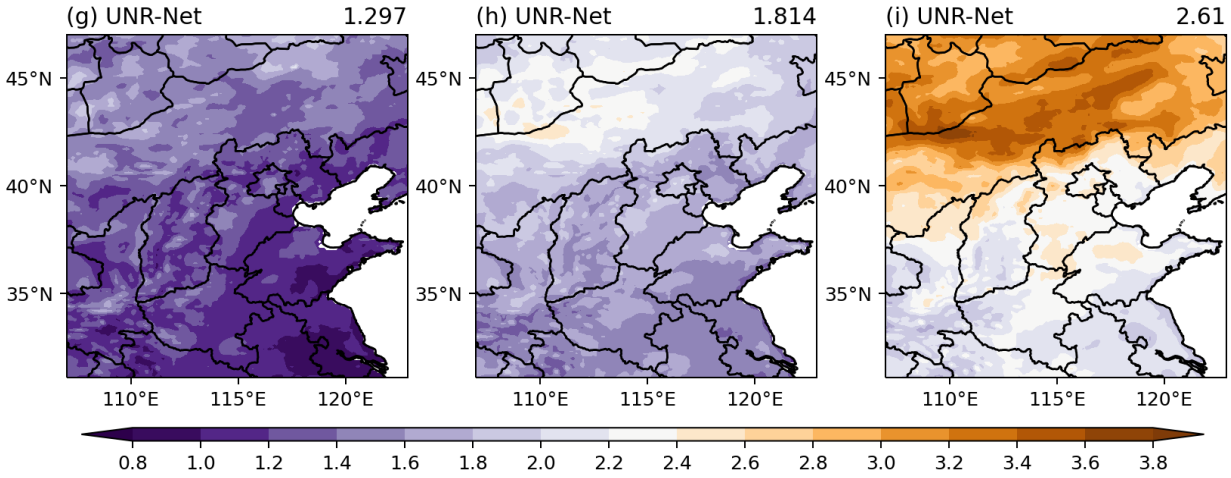


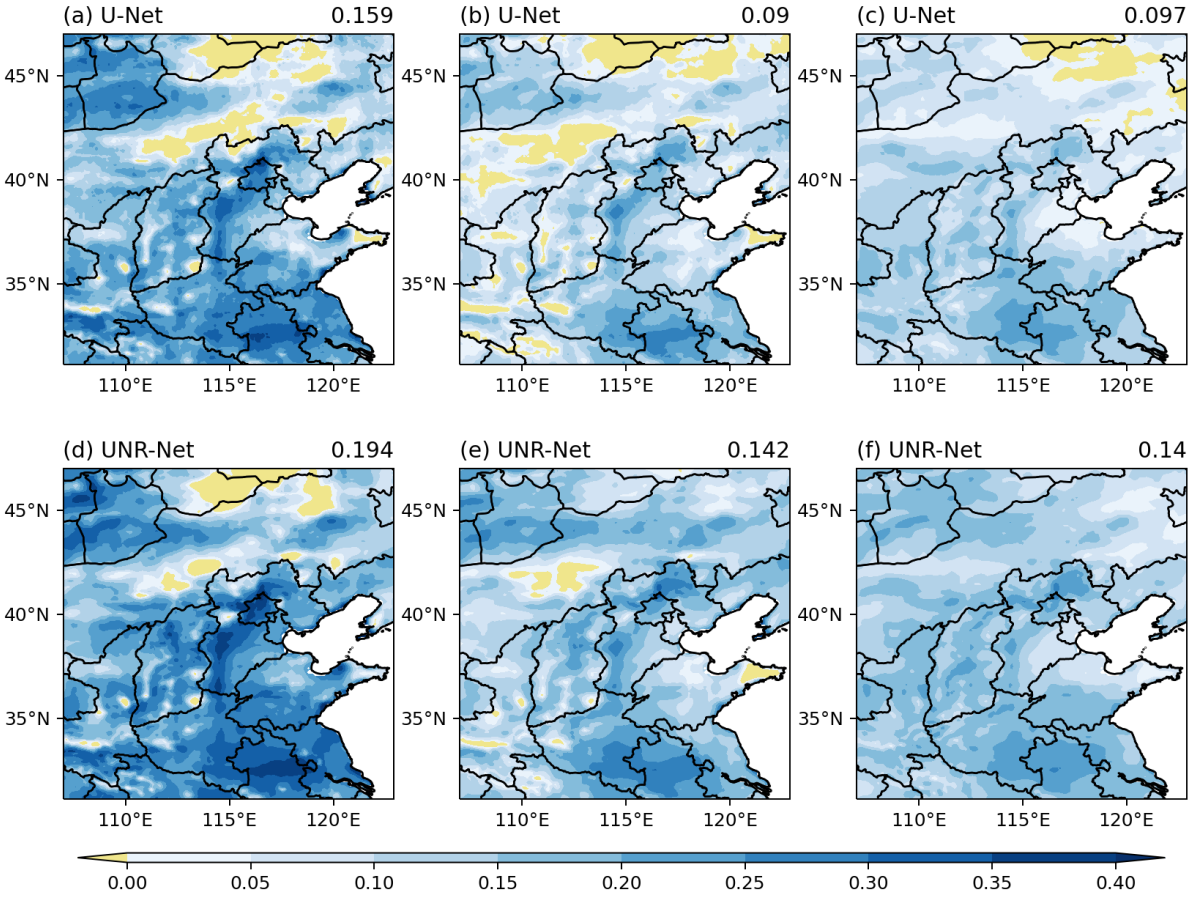
Figure 7. Spatial distributions of the RMSE for 2-m temperature with lead times of 1, 4, and 7 days derived from LR (a–c), U-Net (d–f), and UNR-net (g–i) , the values in the upper-right title represent the mean of RMSE in each case.

The spatial distribution of the root mean square error (RMSE) predominantly demonstrates a pattern of lower values in the southeast and higher values in the northwest (Figure 7). Furthermore, as the lead time increases, the overall error tends to escalate. At a lead time of 1 day, the error distribution of the LR method appears to be relatively uniform. While the region with errors below 0.6 °C is predominantly concentrated in the eastern part, the overall differences are not significant. On the other hand, U-Net and UNR-Net exhibit smaller errors in the southeastern region, with large areas showing errors below 1.2 °C. Moreover, UNR-Net showcases a larger region with errors below 1 °C compared to U-Net. Specifically, UNR-Net demonstrates a greater coverage of areas with errors below 1°C in Jiangsu, Anhui, and even in Shandong, surpassing the performance of U-Net. Additionally, UNR-Net displays reduced errors in regions with complex topography such as the Shandong Peninsula, Shaanxi, and Shanxi, with a higher number of areas exhibiting errors below 1.2°C. This highlights the stronger correction capability of UNR-Net. At a lead time of 4 days, the LR method demonstrates an overall spatial distribution pattern of lower errors in the eastern and western regions, with higher errors observed in the central area. Specifically, there are areas in the central part of Hebei province where errors exceed 2.2°C, and some regions even exhibit errors surpassing 2.6°C. In contrast, the spatial distribution of the two deep learning methods does not exhibit a distinct pattern of lower errors in the central region. Therefore, compared to the LR method, there is a significant

improvement in the central region, resulting in an overall pattern of lower errors in the southern areas and higher errors in the northern areas. Notably, UNR-Net outperforms U-Net in terms of the southern region, with a larger area showing errors below 1.6°C. This is especially evident in the complex terrain areas of southern Shaanxi and central Shanxi, where UNR-Net demonstrates even lower errors, showcasing its superior downscaling capability in complex topography conditions. At a lead time of 7 days, the error distribution of the LR method is similar to that at a 3-day lead time. Both deep learning methods also exhibit a pattern of higher errors in the northern regions and lower errors in the southern regions. Additionally, UNR-Net exhibits lower errors in the regions of Shandong and Hebei, showcasing its superior performance in those areas. Moreover, UNR-Net demonstrates better downscaling capability in the southern regions of Shaanxi and Shanxi as well.

Subsequently, a further comparison was made between the absolute error performance of the two deep learning methods (Figure 8). Overall, when comparing the improvement levels of the two deep learning methods relative to the LR method, UNR-Net exhibits a greater degree of improvement compared to U-Net. The RMSESS scores of UNR-Net, at lead times of 1, 4, and 7 days, were found to be higher than those of U-Net by 0.035, 0.052, and 0.043, respectively. It can be observed that as the lead time increases, UNR-Net demonstrates a greater overall improvement over the LR method. From the spatial distribution, it can be observed that the significant improvements of both deep learning methods are concentrated in central Hebei, Jiangsu, Anhui, and Henan. Across all lead times, these regions exhibit the lowest scores in the entire area. At a lead time of 1 day, there are areas where the score is less than zero, indicating minimal improvement compared to the LR method. These areas are mainly located in Inner Mongolia, with UNR-Net exhibiting a smaller coverage compared to U-Net. In the Shandong region, unlike U-Net, UNR-Net does not have any areas with scores less than 0. Comparing the areas with higher scores, it can be observed that in central Hebei and Anhui, UNR-Net covers a significantly larger area with scores above 0.3, and even areas with scores above 0.35, while U-Net only has a small portion of the western Anhui region with scores above 0.35, and no such distribution in other regions. The areas where significant improvements were observed are primarily concentrated around the Taihang Mountains and the southeastern region. At lead times of 4 and 7 days, the overall situation is similar to that at a lead time of 1 day. UNR-Net outperforms U-Net in terms of higher scores. Particularly at a lead time of seven days, UNR-Net

563 shows improvements over the LR method in all regions. Thus, it further demonstrates the
564 superiority of UNR-Net, particularly at longer lead times.



565

566 **Figure 8.** RMSESS spatial distribution of deep learning methods relative to LR methods for 2-m
567 temperature with lead times of 1, 4, and 7 days derived from U-Net (a–c) and UNR-net (d–f) ,
568 the values in the upper-right title represent the mean of RMSESS in each case.

569 Choosing a day from the testing dataset characterized by a widespread occurrence of low-
570 temperature rain and snow events, we present three illustrative examples of downscaling
571 methods (Figure 9). All three methods aim to downscale the low-resolution forecast data shown
572 in Figure 12a to achieve results that closely resemble the high-resolution observational data
573 depicted in Figure 12b. In this particular case, the forecast data generally exhibit lower values in
574 the northern region compared to the observational data. This highlights the need for higher
575 requirements in terms of correcting the forecast data. From the downscaling results of the three
576 methods, it can be observed that all three methods exhibit finer textures compared to the low-

resolution forecast data. In the case of the LR method, the blue color in the northern region appears lighter, indicating higher temperatures compared to the observed values and the downscaled results of the other two methods. Therefore, the LR method exhibits larger errors. On the other hand, both deep learning methods show significantly smaller errors compared to the LR method, with UNR-Net demonstrating even smaller errors than U-Net.

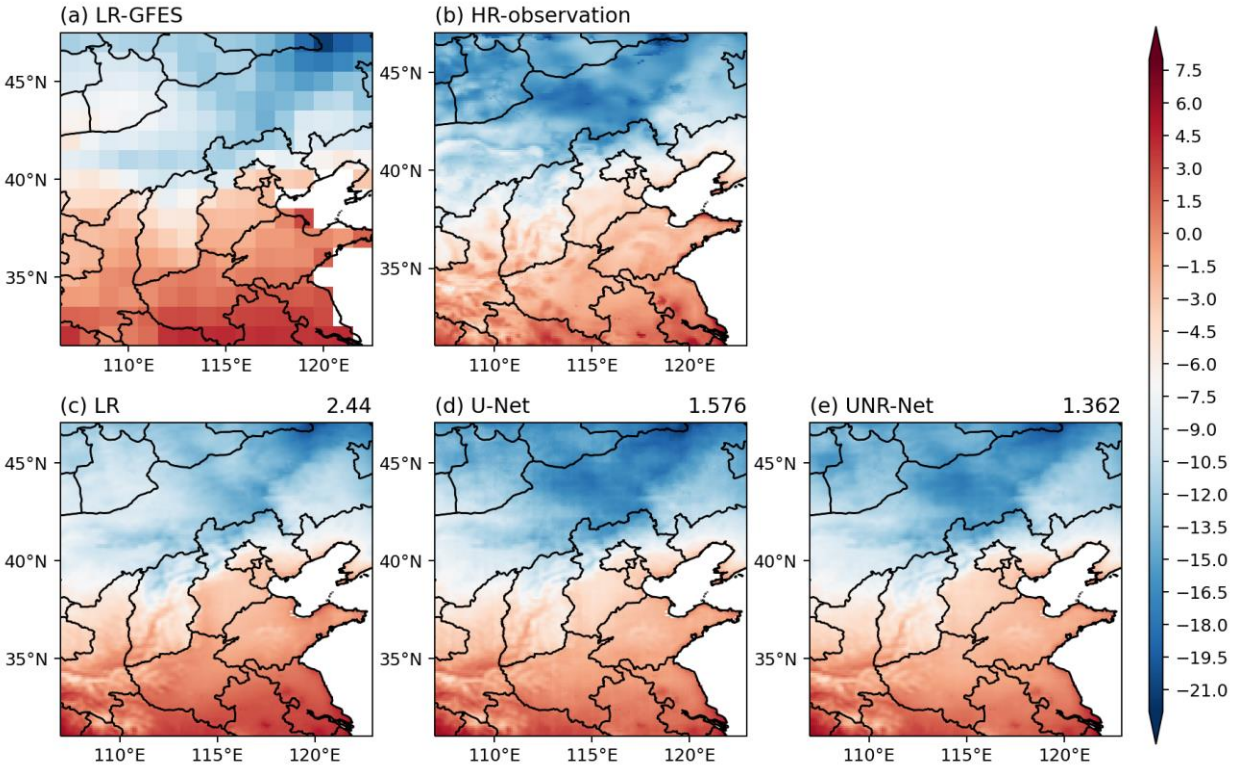


Figure 9. The 2-m temperature downscaling example for a lead time of 1 day on 6 December 2019, with the (a) low-resolution forecast, (b) high-resolution observation, (c) LR downscaling result, (d) U-Net downscaling result and (e) UNR-Net downscaling result. The value in the upper-right title is the MSE of the three downscaled results.

The performance of a method in capturing extreme values is also an important criterion for assessing its effectiveness in the context of low-temperature rain and snow events. From Figure 10, it is evident that for 6 December 2019, the downscaled results of the LR method consistently deviate from the red line. The values tend to be higher overall, indicating lower accuracy. Furthermore, the distribution of points appears to be widely scattered especially in the low-temperature range. The two deep learning methods exhibit significant improvements in accuracy compared to LR. The extent of deviation from the red line is reduced, and the points generally

align along the line. Additionally, the distribution of points is much more concentrated compared to the LR method. For the two deep learning methods, UNR-Net demonstrates a better performance than U-Net in the low-temperature range. The points are more concentrated, with the distribution center aligning closely with the red line.

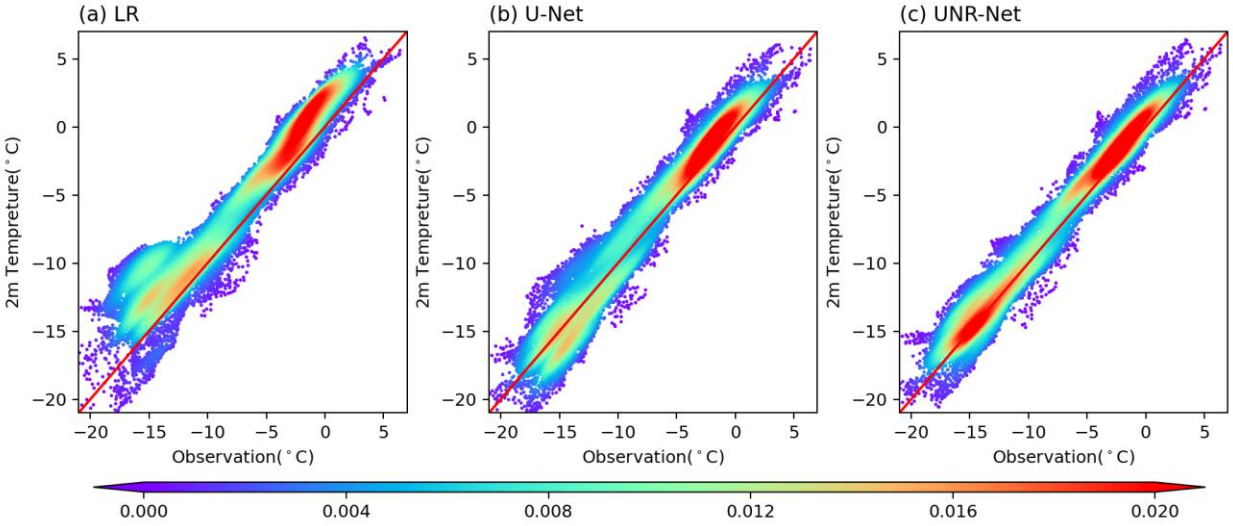


Figure 10. Scatter plot density of the downscaled results for the three methods on 6 December 2019, with a 24-h lead time derived from (a)LR, (b) U-Net, and (c) UNR-net.

As illustrated in the above example, for the majority of days in the testing dataset, the performance of the three methods is characterized by LR being the poorest and UNR-Net being the best (Figure 11). However, when considering the performance over the entire year in the testing dataset, noticeable seasonal variations can be observed. Regarding NSE and PCC, the performance of the low-resolution forecast data is not stable. It exhibits significant fluctuations and lower scores, particularly around the summer season. However, after downscaling using the three methods, the scores of the results remain stable throughout the year. This indicates a significant improvement in the fitting and correlation between the predicted and observed fields, particularly during the summer season. In terms of RMSE, although the error of the low-resolution forecast data fluctuates significantly at the beginning and end of the year and remains relatively constant throughout the middle of the year, the overall numerical value remains consistent. However, the error values of the downscaled results using the three methods do not exhibit a constant distribution. They generally show a pattern of being lower during the middle of the year and higher at the beginning and end of the year. However, the fluctuations are

relatively evenly distributed throughout the year. In terms of SSIM, the low-resolution forecast data generally exhibits slightly higher scores during the middle of the year compared to the beginning and end of the year, with less pronounced fluctuations. However, there are significant differences in the distribution of results for the three downscaled methods. The scores are noticeably higher and more stable during the middle of the year compared to the beginning and end of the year. Hence, it is necessary to conduct further analysis based on seasons to gain deeper insights.

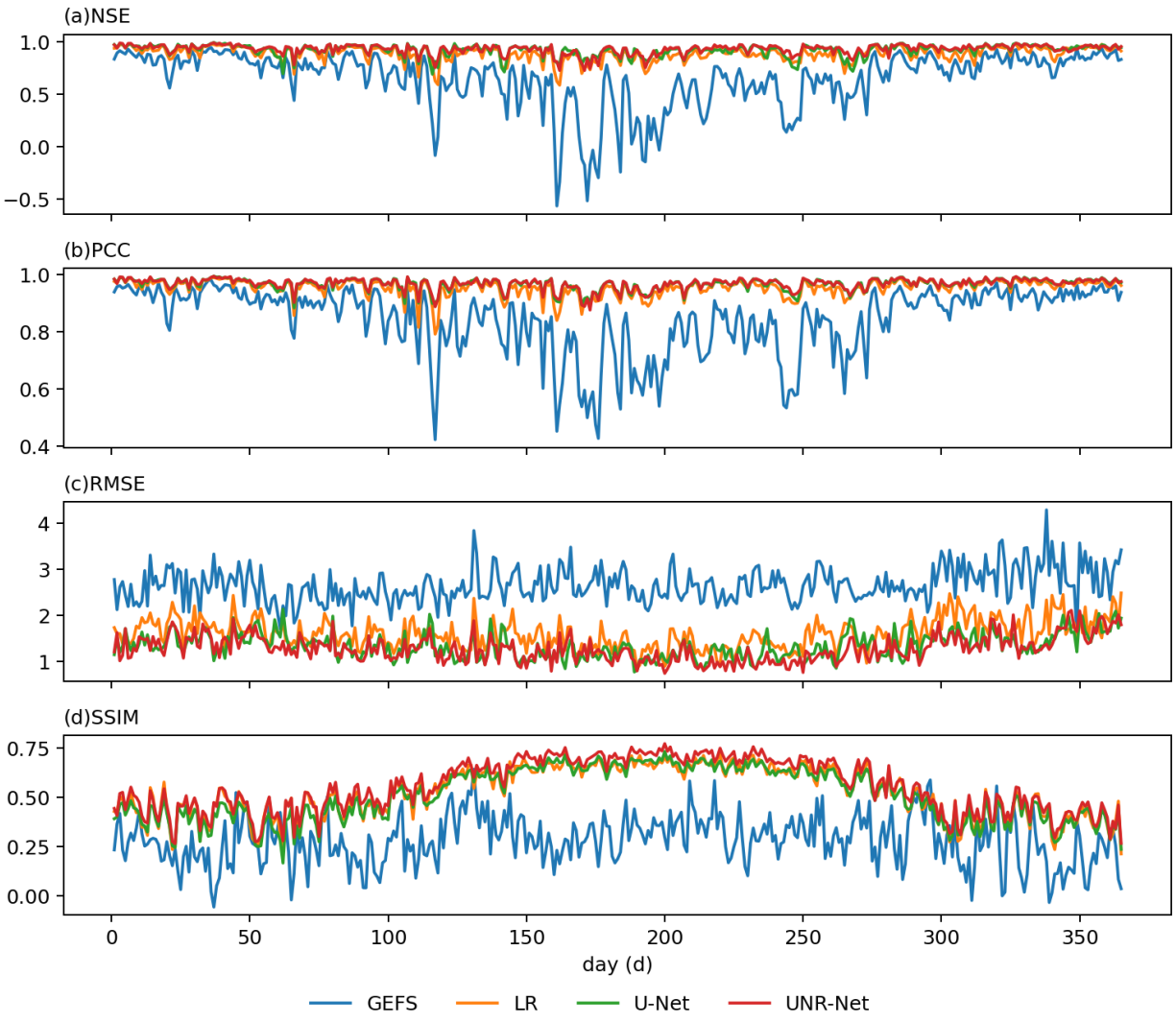


Figure 11. The error performance of the downscaled results for the three methods in the whole testing dataset and the GEFS data with a 1-d lead time for (a) NSE, (b) PCC, (c) RMSE, and (d) SSIM.

4.2 Error decomposition

The decomposition of MSE into four seasons (spring, summer, autumn, and winter) for lead times of 1–7 days is illustrated in Figure 12. For the low-resolution forecast data, as the lead time increases, the error also increases. However, the rate of increase is the smallest during the summer season. The difference in error between a lead time of 7 days and 1 day is 1.10 in the summer season, while it is 2.10 in the spring season, 1.58 in the autumn season, and 1.50 in the winter season. Moreover, as the lead time increases, the error of the LR method becomes closer to the error of the low-resolution forecast data in all four seasons. The difference between the two decreases by 0.13, 0.22, 0.06, and 0.11 in the spring, summer, autumn, and winter seasons, respectively, when comparing a lead time of 7 days to a lead time of 1 day. Particularly, the error in the autumn and winter seasons shows a closer growth rate to that of the low-resolution forecast data. As for the two deep learning methods, the difference between their errors and the errors of the low-resolution forecast data changes differently compared to the performance of the LR method as the lead time increases. For the U-Net method, the difference between the two decreases by 0.035 and 0.017 in the autumn and winter seasons, respectively, when comparing a lead time of 7 days to a lead time of 1 day. However, in the spring and summer seasons, the difference increases by 0.20 and 0.10, respectively. On the other hand, for the UNR-Net method, the difference between its error and the error of the low-resolution forecast data increases in all four seasons as the lead time increases. The difference at a lead time of 7 days compared to a lead time of 1 day increases by 0.45, 0.20, 0.08, and 0.20 in the spring, summer, autumn, and winter seasons, respectively. This not only highlights the advantages of the two deep learning methods, particularly UNR-Net, over the LR method but also further emphasizes that as the lead time increases, the advantages of the deep learning methods, especially UNR-Net, become more significant and comprehensive. Additionally, both deep learning methods demonstrate greater advantages in the spring and summer seasons.

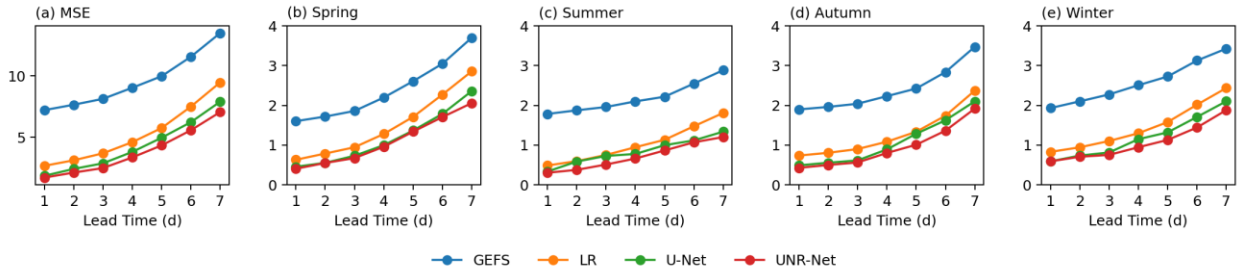


Figure 12. The decomposition of errors averaged over North China for each season across lead times of 1 to 7 days, with (a) MSE, (b) spring, (c) summer, (d) autumn, and (e) winter.

As shown in Figure 13, when decomposing the errors for the four seasons, it can be observed that the three downscaled methods exhibit significant improvements in the Bias and Distribution components. For the Bias term, the average values of the LR method across the four seasons are 0.02, 0.03, 0.02, and 0.06, respectively. The average values of the U-Net method across the four seasons are 0.009, 0.03, 0.01, and 0.02, respectively. As for UNR-Net, the average values across the four seasons are 0.007, 0.007, 0.009, and 0.006, respectively. Indeed, it can be observed that UNR-Net has the smallest Bias term, with values below 0.01 in all four seasons. Particularly in the winter season, the difference between the LR method and the two deep learning methods is much more significant compared to the other three seasons. For the Distribution term, the average values of the LR method across the four seasons are 0.212, 0.124, 0.297, and 0.328, respectively. The average values of the U-Net method across the four seasons are 0.171, 0.215, 0.246, and 0.182, respectively. As for UNR-Net, the average values across the four seasons are 0.164, 0.124, 0.197, and 0.165, respectively. The patterns for the Distribution term are similar to those of the Bias term. UNR-Net consistently exhibits the lowest error, and in the winter season, the difference between the LR method and the two deep learning methods is significantly larger compared to the other three seasons.

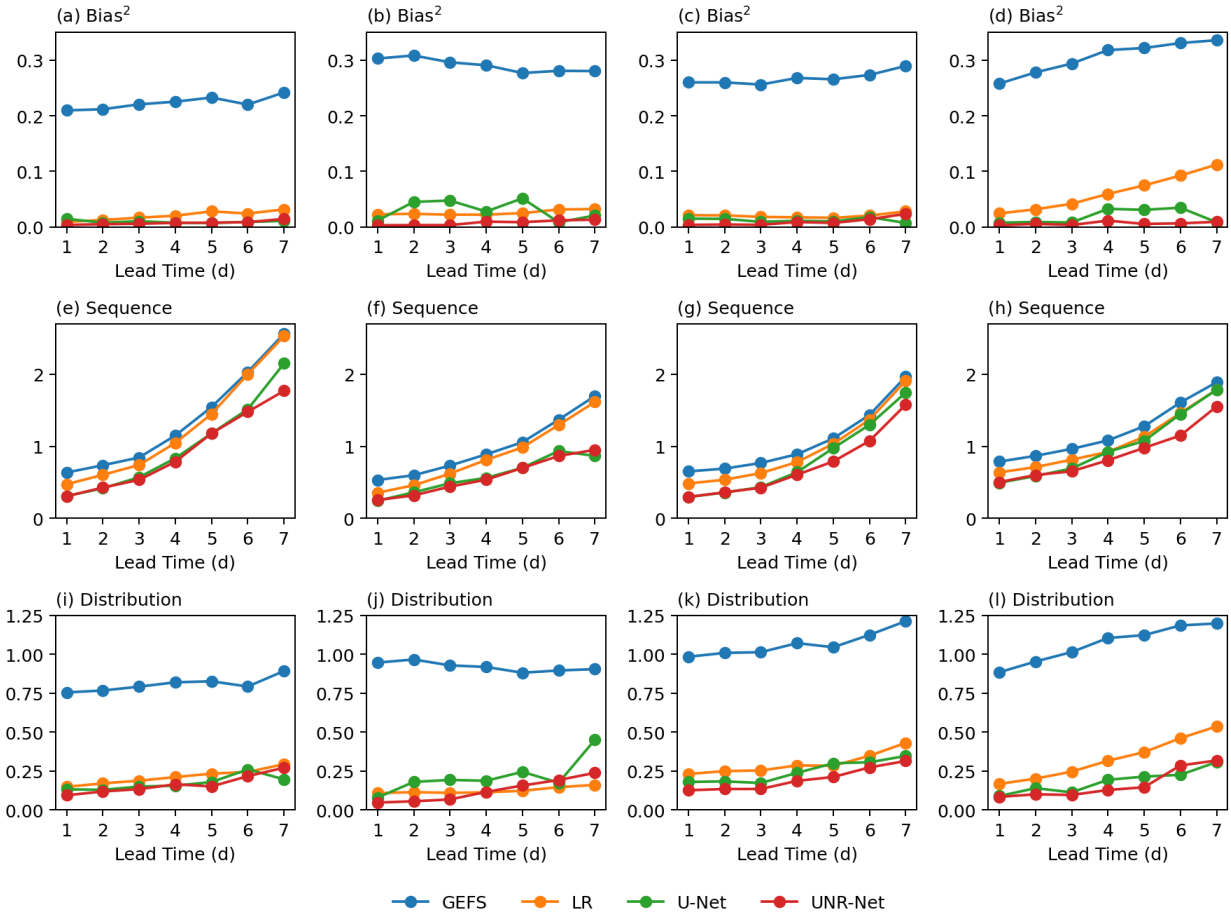


Figure 13. The values of Bias (a–d), Sequence (e–h), and Distribution (i–l) error components decomposed for lead times of 1 to 7 days for each season, (a, e, i) spring, (b, f, j) summer, (c, g, k) autumn, and (d, h, l) winter, averaged over North China.

As for the Sequence term, the LR method shows limited capability, especially at longer lead times, where the improvement relative to the low-resolution forecast data is minimal. At a lead time of 7 days, the difference between the low-resolution forecast and the LR method is only 0.034, 0.083, 0.057, and 0.103 across the four seasons, respectively. On the other hand, at a lead time of 1 day, the difference between the two is 0.163, 0.177, 0.167, and 0.150 in the respective seasons. Clearly, as the lead time increases, the correction capability of the LR method becomes weaker. Indeed, both deep learning methods demonstrate advantages over the LR method. The discrepancies between the U-Net method and the low-resolution forecasts for the four seasons are as follows: at a lead time of 1 day, they are 0.327, 0.289, 0.356, and 0.294, and at a lead time of 7 days, they are 0.414, 0.832, 0.221, and 0.104. It can be observed that as the lead time

increases, the accuracy improvement of U-Net becomes more prominent during the spring and summer seasons. The disparities between the UNR-Net method and the low-resolution forecasts for the four seasons are as follows: at a lead time of 1 day, they are 0.329, 0.279, 0.354, and 0.284, and at a lead time of 7 days, they are 0.794, 0.748, 0.389, and 0.337. It can be observed that as the lead time increases, the differences between the downscaled results of UNR-Net and the low-resolution forecasts intensify across all four seasons. This indicates that the UNR-Net method exhibits a greater degree of improvement over low-resolution forecasts with longer lead times. This observation indicates that the Sequence component highlights the advantages of nonlinear methods to a greater extent. Deep learning methods primarily improve the accuracy of downscaling tasks in the temporal domain.

The proportion of errors for each component after error decomposition is illustrated in Figure 14. For the errors associated with the Bias, Sequence, and Distribution components, Bias has the smallest proportion. The average proportions for the LR method, U-Net, and UNR-Net across the four seasons are 0.59%, 0.42%, and 0.21% for Bias, respectively. Next, is the Distribution component, with average proportions of 4.86%, 5.04%, and 4.43% for the LR method, U-Net, and UNR-Net across the four seasons, respectively. The dominant component is Sequence, with average proportions of 19.55%, 19.54%, and 20.36% for the LR method, U-Net, and UNR-Net across the four seasons, respectively. Therefore, the Sequence component plays a more significant role in determining the performance of the methods.

As the lead time increases, the proportion of the Sequence component gradually increases. For a lead time of 1 day, the average proportions of the three methods across the four seasons are 18.30%, 18.04%, and 19.74% respectively. For a lead time of 4 days, the average proportions are 19.51%, 19.56%, and 20.44% respectively. For a lead time of 7 days, the average proportions are 20.84%, 21.02%, and 20.91% respectively. This trend may be attributed to the fact that the errors in the forecast data in terms of temporal variability increase with longer lead times, resulting in a higher proportion of temporal errors in the downscaled results of the three methods.
Furthermore, there have been changes in the proportions across seasons. For a lead time of 1 day, the average proportions of the three methods across the four seasons are 7.98%, 6.17%, 8.59%, and 10.58% respectively. It can be observed that the majority of errors are concentrated in the winter season. When the lead time increases to 4 days, the average proportions across the four

seasons are 9.43%, 6.76%, 7.75%, and 9.40% respectively. For a lead time of 7 days, the average proportions across the four seasons are 10.37%, 5.98%, 8.54%, and 8.44%, respectively. It can be noted that with the increase in lead time, the seasons with higher proportions of errors gradually shift toward the spring season.

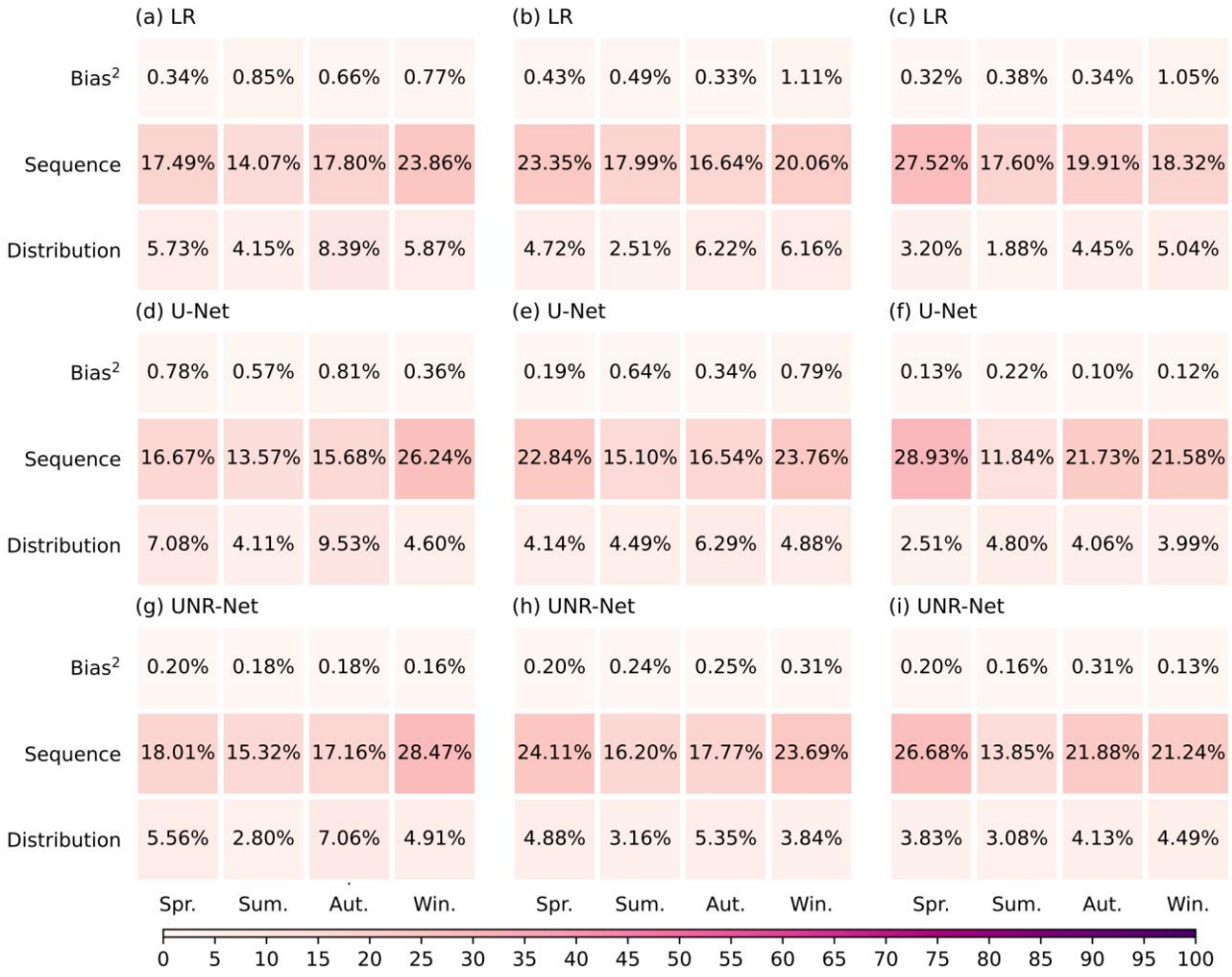


Figure 14. The decomposed 12 error components averaged over North China, represented as percentages with lead times of (a, d, g) 1, (b, e, h) 4, and (c, f, i) 7 days, derived from (a–c) LR, (d–f) U-Net, and (g–i) UNR-Net. The vertical axis represents errors for physical significance and the horizontal axis represents errors for the four seasons.

The numerical values for the decomposed error components are depicted in Figure 15. Based on the numerical values of the 12 error components, as the lead time increases, for lead times of 1, 4, and 7 days, the average values of the Bias component for the three methods across the four seasons are 0.011, 0.019, and 0.013, respectively. The average values of the Sequence

component are 0.388, 0.771, and 1.551, respectively. The average values of the Distribution component are 0.123, 0.191, and 0.305, respectively. It can be observed that the Sequence and Distribution components show significant increases, while the Bias component remains relatively stable. Furthermore, across all components and lead times, U-Net outperforms LR, and UNR-Net outperforms U-Net.

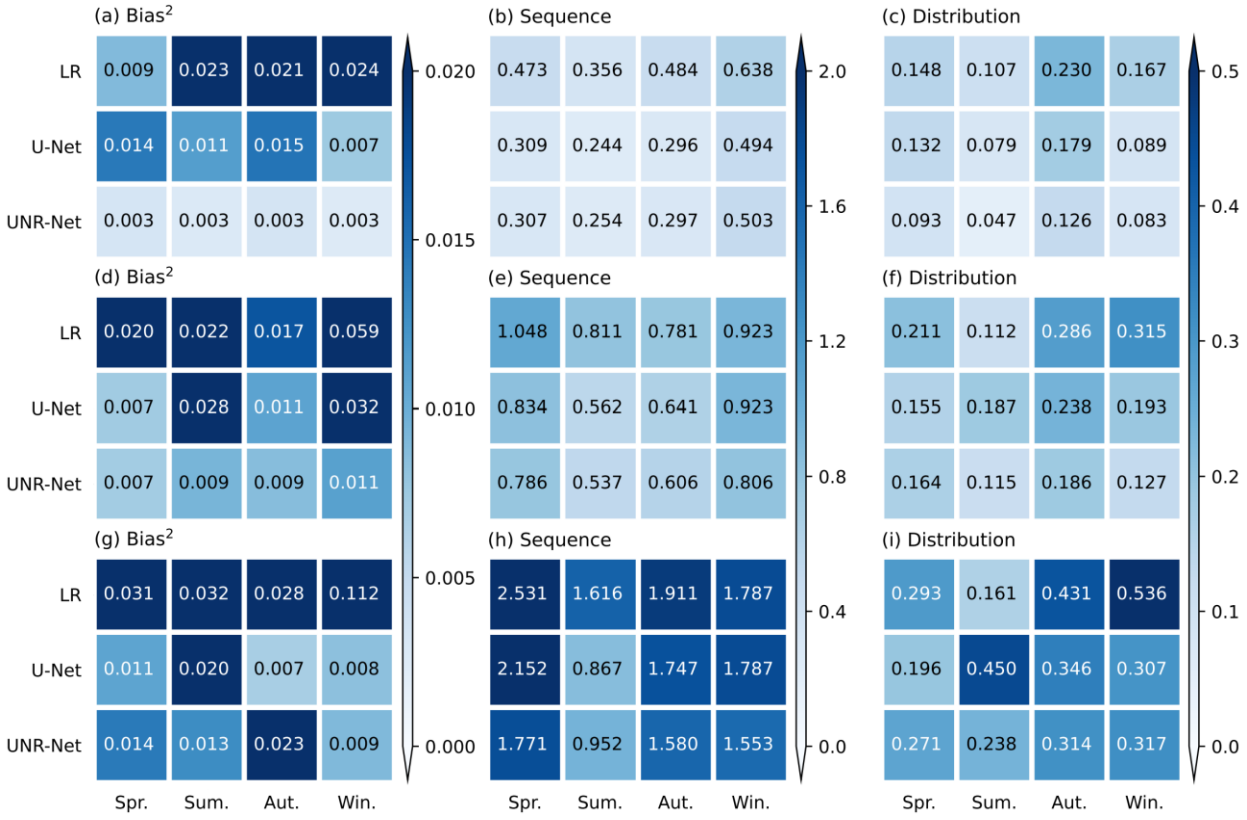


Figure 15. The decomposed error values of the 12 components averaged over North China with lead times of (a–c) 1, (d–f) 4, and (g–i) 7 days. The vertical axis represents errors from three different methods and the horizontal axis represents the seasonal error.

For different seasons, the average values of the Bias component across the three lead times are 0.013, 0.017, 0.014, and 0.030, respectively. Although the Bias component is almost twice as large in the winter season compared to the other three seasons, its contribution to the overall MSE is relatively small. Therefore, the winter season does not exhibit significantly higher errors compared to the other three seasons due to this component. The average values of the Distribution component across the three lead times are 0.185, 0.166, 0.260, and 0.237, respectively. The errors in the spring and summer seasons are smaller than the other two seasons.

The average values of the Sequence component across the three lead times are 1.135, 0.689, 0.927, and 1.046, respectively. The Sequence component exhibits much better performance in summer compared to the other three seasons. Moreover, the Sequence component has the highest values numerically, indicating its dominant role in MSE.

5. Conclusions and discussion

This paper introduces a novel downscaling network called UNR-Net, which integrates a non-local attention mechanism, Res2net (Gao et al., 2019), and terrain information to further enhance the accuracy and practical value of the results. A downscaling experiment with a downscaling factor of 10x was conducted for the 2-m temperature forecast over the East China region at lead times of 1–7 days. The LR and U-Net methods are conducted as benchmarks. To obtain a more detailed and specific evaluation and enhance the interpretability of the models, the error decomposition method based on MSE is also proposed.

Generally, the UNR-Net demonstrates superior performance over U-Net and LR methods in terms of NSE, PCC, RMSE, and SSIM, particularly for longer lead times. Regarding NSE, PCC, and RMSE, the LR method exhibits the poorest performance, followed by U-Net. The best-performing method is UNR-Net. Both deep learning methods demonstrated a certain improvement compared to the LR method when forecasting for longer lead times. Moreover, UNR-Net exhibited a more pronounced enhancement compared to U-Net. For SSIM, the U-Net method shows the poorest performance, followed by the LR method, while UNR-Net exhibits the best performance. Therefore, it can be observed that UNR-Net has superior practical applicability compared to U-Net. In terms of spatial distribution, the errors are primarily concentrated in regions with complex terrain, such as the Taihang Mountains, Shanxi, central Shaanxi, and Liaoning. UNR-Net exhibits significantly smaller errors in this area compared to the other two methods, indicating its greater advantage in complex terrain regions. Furthermore, it was observed that during the summer season, characterized by lower NSE and PCC values in the low-resolution data, all three methods exhibited better performance in terms of RMSE and SSIM.

Consequently, for a more in-depth analysis of the errors, the Mean Squared Error (MSE) is first decomposed based on time into four seasons: spring, summer, autumn, and winter. Then, it is further decomposed based on its physical significance into three components: Bias, Sequence,

and Distribution. Each method's error is decomposed into 12 constituent components. Indeed, it can be observed that the three methods showed the lowest errors during the summer season. Moreover, the deep learning methods, especially UNR-Net, displayed more significant advantages as the lead time increased. Upon decomposing the errors for each season into Bias, Sequence, and Distribution components, it can be observed that for the Bias and Distribution components, all three methods showed significant improvements in downscaling results compared to low-resolution data, with UNR-Net exhibiting the smallest error. Among the error composition components, the Sequence component has the largest proportion and plays a dominant role. Especially for longer lead times, the LR method showed little improvement compared to low-resolution data, while both deep learning methods demonstrated higher accuracy, with UNR-Net showing the smallest errors.

The success of UNR-Net in temperature downscaling highlights the feasibility of utilizing deep learning methods and techniques such as non-local attention mechanisms and residual connections for handling Earth system data. Although UNR-Net has already incorporated terrain data, it lacks the utilization of additional meteorological variables. Existing studies have shown that the integration of diverse meteorological variables can enhance the accuracy of results (Sun & Tang, 2020; Harris et al., 2022). Therefore, in the future, it is worth considering the incorporation of more meteorological elements into the downscaling task to further improve its performance. On the other hand, with the ongoing advancements in deep learning technology, there exists significant potential for further improvements in result accuracy and exploration of new possibilities. Moreover, from an analysis of error decomposition, it is evident that the degree of improvement varies for different error components. Therefore, in the future, it would be beneficial to consider employing techniques tailored to specific physical meanings or seasons. Incorporating approaches that target seasonality, mean values, temporal patterns, and distributions, such as season-based transfer learning, holds the potential to not only enhance overall error performance but also increase their practical value significantly.

Acknowledgments

This paper is supported by the National Natural Science Foundation of China (42275164) and the Science and Technology Program of China Southern Power Grid Co., Ltd. (Grant No.

YNKJXM202222172), the Reserve Talents Program for Middle-aged and Young Leaders of Disciplines in Science and Technology of Yunnan Province, China (Grant No. 202105AC160014). We are grateful to ECMWF and NCEP/NOAA for their datasets.

Data Availability Statement

Data related to this article are available free from Global Ensemble Forecasting System (https://noaa-gefs-retrospective.s3.amazonaws.com/index.html#GEFSv12/reforecast/) (Guan et al., 2020), ERA5-Land (https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-land?tab=form) (Muñoz-Sabater et al, 2021) and ETOPO1 (https://www.ngdc.noaa.gov/thredds/catalog/global/ETOPO2022/60s/60s_bed_elev_netcdf/catalog.html?dataset=globalDatasetScan/ETOPO2022/60s/60s_bed_elev_netcdf/ETOPO_2022_v1_60s_N90W180_bed.nc) (Amante & Eakins, 2009). These data have been processed with Python (version 3.8.8). The training process was executed using NVIDIA RTX A5000 under PyTorch1.11, which can be accessed from https://pytorch.org/.

References

Amante, C., & Eakins, B. (2009). ETOPO1 arc-minute global relief model: Procedures, data sources and analysis [Dataset]. *NOAA Tech. Memo*, NESDIS NGDC-24, 25 pp.

Baño-Medina, J., Manzanas, R., & Gutiérrez, J. M. (2020). Configuration and intercomparison of deep learning neural models for statistical downscaling. *Geoscientific Model Development, 13*(4), 2109–2124.

Chen, J., Chen, H., & Guo, S. (2018). Multi-site precipitation downscaling using a stochastic weather generator. *Climate Dynamics, 50*, 1975–1992.

824 Dong, C., Loy, C. C., He, K., & Tang, X. (2015). Image super-resolution using deep
825 convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 38*(2), 295–307.
826
827 Doury, A., Somot, S., Gadat, S., Ribes, A., & Corre, L. (2023). Regional Climate Model
828 emulator based on deep learning: concept and first evaluation of a novel hybrid downscaling
829 approach. *Climate Dynamics, 60*(56), 1751–1779.
830 Dumoulin, V., Belghazi, I., Poole, B., Mastropietro, O., Lamb, A., Arjovsky, M., & Courville, A.
831 (2016). Adversarially learned inference. *arXiv preprint arXiv:1606.00704.*
832 Feser, F., Rockel, B., von Storch, H., Winterfeldt, J., & Zahn, M. (2011). Regional climate
833 models add value to global model data: a review and selected examples. *Bulletin of the American*
834 *Meteorological Society, 92*(9), 1181–1192.
835 Fowler, H. J., Blenkinsop, S., & Tebaldi, C. (2007). Linking climate change modelling to
836 impacts studies: recent advances in downscaling techniques for hydrological modelling.
837 *International Journal of Climatology: A Journal of the Royal Meteorological Society, 27*(12),
838 1547–1578.
839 Frei, C., Christensen, J. H., Déqué, M., Jacob, D., Jones, R. G., & Vidale, P. L. (2003). Daily
840 precipitation statistics in regional climate models: Evaluation and intercomparison for the
841 European Alps. *Journal of Geophysical Research: Atmospheres, 108*(D3).
842 Gao, S. H., Cheng, M. M., Zhao, K., Zhang, X. Y., Yang, M. H., & Torr, P. (2019). Res2net: A
843 new multi-scale backbone architecture. *IEEE transactions on pattern analysis and machine*
844 *intelligence, 43*(2), 652–662.

Gauthier, J. (2014). Conditional generative adversarial nets for convolutional face generation. *Class project for Stanford CS231N: convolutional neural networks for visual recognition, Winter semester, 2014*(5), 2.

Gerges, F., Boufadel, M. C., Bou-Zeid, E., Nassif, H., & Wang, J. T. (2022, January). A novel deep learning approach to the statistical downscaling of temperatures for monitoring climate change. In *2022 The 6th International Conference on Machine Learning and Soft Computing* (pp. 1–7).

Guan H, Zhu Y, Sinsky E, et al. (2020). The NCEP GEFS-v12 reforecasts to support subseasonal and hydrometeorological applications [Dataset]. In *Climate Prediction S&T Digest, 44rd NOAA Climate Diagnostics and Prediction Workshop special issue*.

Hagemann, S., Machenhauer, B., Jones, R., Christensen, O. B., Déqué, M., Jacob, D., & Vidale, P. L. (2004). Evaluation of water and energy budgets in regional climate models applied over Europe. *Climate Dynamics, 23*, 547–567.

Han L, Chen M, Chen K, et al. (2021). A deep learning method for bias correction of ECMWF 24–240 h forecasts. *Advances in Atmospheric Sciences, 38*(9), 1444–1459.

Harris, L., McRae, A. T., Chantry, M., Dueben, P. D., & Palmer, T. N. (2022). A generative deep learning approach to stochastic downscaling of precipitation forecasts. *Journal of Advances in Modeling Earth Systems, 14*(10), e2022MS003120.

He, K., Zhang, X., Ren, S., & Sun, J. (2016a). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

He, X., Chaney, N. W., Schleiss, M., & Sheffield, J. (2016b). Spatial downscaling of precipitation using adaptable random forests. *Water Resources Research, 52*(10), 8217–8237.

867 Hodson, T. O., Over, T. M., & Foks, S. S. (2021). Mean squared error, deconstructed. *Journal of*
868 *Advances in Modeling Earth Systems, 13*(12), e2021MS002681.
869 Höhlein, K., Kern, M., Hewson, T., & Westermann, R. (2020). A comparative study of
870 convolutional neural network models for wind field downscaling. *Meteorological Applications,*
871 *27*(6), e1961.
872 Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE*
873 *conference on computer vision and pattern recognition* (pp. 7132–7141).
874 Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected
875 convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern*
876 *recognition* (pp. 4700–4708).
877 Ioffe, S., & Szegedy, C. (2015, June). Batch normalization: Accelerating deep network training
878 by reducing internal covariate shift. In *International conference on machine learning* (pp. 448–
879 456). pmlr.
880 Ji, Y., Gong, B., Langguth, M., Mozaffari, A., & Zhi, X. (2022). CLGAN: A GAN-based video
881 prediction model for precipitation nowcasting. *EGUsphere*, 1–23.
882 Ji, L., Zhi, X., Schalge, B., Stephan, K., Wu, Z., Wu, C., … & Zhu, S. (2023a). Dynamic
883 downscaling ensemble forecast of an extreme rainstorm event in South China by COSMO EPS.
884 *AI-based prediction of high-impact weather and climate extremes under global warming: A*
885 *perspective from the large-scale circulations and teleconnections, 16648714*, 143.
886 Ji, Y., Zhi, X., Ji, L., Zhang, Y., Hao, C., & Peng, T. (2023b). Deep-learning-based. *AI-based*
887 *prediction of high-impact weather and climate extremes under global warming: A perspective*
888 *from the large-scale circulations and teleconnections, 16648714*, 200.

Jing, Y., Lin, L., Li, X., Li, T., & Shen, H. (2022). An attention mechanism based convolutional network for satellite precipitation downscaling over China. *Journal of Hydrology, 613*, 128388.

Jones, P. W. (1999). First-and second-order conservative remapping schemes for grids in spherical coordinates. *Monthly Weather Review, 127*(9), 2204–2210.

Kim, G., & Barros, A. P. (2002). Downscaling of remotely sensed soil moisture with a modified fractal interpolation method using contraction mapping and ancillary data. *Remote Sensing of Environment, 83*(3), 400–413.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv*:1412.6980.

Kumar, B., Chattopadhyay, R., Singh, M., Chaudhari, N., Kodari, K., & Barve, A. (2021). Deep learning–based downscaling of summer monsoon rainfall data over Indian region. *Theoretical and Applied Climatology, 143*, 1145–1156.

Li, X., Chen, S., Hu, X., & Yang, J. (2019). Understanding the disharmony between dropout and batch normalization by variance shift. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2682–2690).

Mannig B, Müller M, Starke E, et al. (2013). Dynamical downscaling of climate change in Central Asia. *Global and Planetary Change, 110*, 26–39.

Mnih, V., Heess, N., & Graves, A. (2014). Recurrent models of visual attention. *Advances in Neural Information Processing Systems, 27*.

Muñoz-Sabater J, Dutra E, Agustí-Panareda A, et al. (2021). ERA5-Land: A state-of-the-art global reanalysis dataset for land applications [Dataset]. *Earth system science data, 13*(9), 4349-4383.

911 Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I—
912 A discussion of principles. *Journal of Hydrology, 10*(3), 282–290.
913 Pan, B., Hsu, K., AghaKouchak, A., & Sorooshian, S. (2019). Improving precipitation estimation
914 using convolutional neural network. *Water Resources Research, 55*(3), 2301–2321.
915 Park, S., Singh, K., Nellikkattil, A., Zeller, E., Mai, T. D., & Cha, M. (2022, August).
916 Downscaling Earth System Models with Deep Learning. In *Proceedings of the 28th ACM*
917 *SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 3733–3742).
918 Rind, D., Rosenzweig, C., & Goldberg, R. (1992). Modelling the hydrological cycle in
919 assessments of climate change. *Nature, 358*(6382), 119–122.
920 Roberts, C. D., Senan, R., Molteni, F., Boussetta, S., Mayer, M., & Keeley, S. P. (2018). Climate
921 model configurations of the ECMWF Integrated Forecasting System (ECMWF-IFS cycle 43r1)
922 for HighResMIP. *Geoscientific Model Development, 11*(9), 3681–3712.
923 Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical
924 image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–*
925 *MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015,*
926 *Proceedings, Part III 18* (pp. 234–241).
927 Sha, Y., Gagne II, D. J., West, G., & Stull, R. (2020). Deep-learning-based gridded downscaling
928 of surface meteorological variables in complex terrain. Part I: Daily maximum and minimum 2-
929 m temperature. *Journal of Applied Meteorology and Climatology, 59*(12), 2057–2073.
930 Sharifi, E., Saghafian, B., & Steinacker, R. (2019). Downscaling satellite precipitation estimates
931 with multiple linear regression, artificial neural networks, and spline interpolation techniques.
932 *Journal of Geophysical Research: Atmospheres, 124*(2), 789–805.

933 Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., … & Hassabis,
934 D. (2017). Mastering the game of go without human knowledge. *Nature, 550*(7676), 354–359.
935 Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a
936 simple way to prevent neural networks from overfitting. *The Journal Of Machine Learning*
937 *Research, 15*(1), 1929–1958.
938 Sun, A. Y., & Tang, G. (2020). Downscaling satellite and reanalysis precipitation products using
939 attention-based deep convolutional neural nets. *Frontiers in Water, 2*, 536743.
940 Vandal, T., Kodra, E., & Ganguly, A. R. (2019). Intercomparison of machine learning methods
941 for statistical downscaling: the case of daily and extreme precipitation. *Theoretical and Applied*
942 *Climatology, 137*, 557–570.
943 Veit, A., Wilber, M. J., & Belongie, S. (2016). Residual networks behave like ensembles of
944 relatively shallow networks. *Advances in Neural Information Processing Systems, 29*.
945 Wang, X., Girshick, R., Gupta, A., & He, K. (2018). Non-local neural networks. In *Proceedings*
946 *of the IEEE conference on computer vision and pattern recognition* (pp. 7794–7803).
947 Wang, F., Tian, D., Lowe, L., Kalin, L., & Lehrter, J. (2021). Deep learning for daily
948 precipitation and temperature downscaling. *Water Resources Research, 57*(4), e2020WR029308.
949 Wilby, R. L., & Wigley, T. M. (1997). Downscaling general circulation model output: a review
950 of methods and limitations. *Progress in Physical Geography*, *21*(4), 530–548.
951 Wilby, R. L., Wigley, T. M. L., Conway, D., Jones, P. D., Hewitson, B. C., Main, J., & Wilks, D.
952 S. (1998). Statistical downscaling of general circulation model output: A comparison of methods.
953 *Water Resources Research, 34*(11), 2995–3008.
954 Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018). Cbam: Convolutional block attention
955 module. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 3–19).

Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1492–1500).

Yuan, Q., Shen, H., Li, T., Li, Z., Li, S., Jiang, Y., … & Zhang, L. (2020). Deep learning in environmental remote sensing: Achievements and challenges. *Remote Sensing of Environment*, *241*, 111716.

Zhu, Y., Zhi, X., Lyu, Y., Zhu, S., Tong, H., Ali, M., … & Huo, W. (2022). Forecast Calibrations of Surface Air Temperature over Xinjiang Based on U-net Neural Network. *Frontiers in Environmental Science*, 1826.