atmosphere

MDPI

Article

Grid-to-Point Deep-Learning Error Correction for the Surface Weather Forecasts of a Fine-Scale Numerical Weather Prediction System

Yu Qin [1], Yubao Liu [1,*], Xinyu Jiang [2], Li Yang [1], Haixiang Xu [3,*], Yueqin Shi [4] and Zhaoyang Huo [1]

[1] Precision Regional Earth Modeling and Information Center, Nanjing University of Information Science and Technology, Nanjing 210044, China
[2] Nanjing Xinda Institute of Meteorological Science and Technology, Nanjing 210044, China
[3] Jibei Electric Power Corporation of the State Grid Corporation of China, Beijing 100031, China
[4] Key Laboratory for Cloud Physics of China Meteorological Administration, Beijing 100081, China
* Correspondence: ybliu@nuist.edu.cn (Y.L.); xhx0617@foxmail.com (H.X.)

Abstract: Forecasts of numerical weather prediction models unavoidably contain errors, and it is a common practice to post-process the model output and correct the error for the proper use of the forecasts. This study develops a grid-to-multipoint (G2N) model output error correction scheme which extracts model spatial features and corrects multistation forecasts simultaneously. The model was tested for an operational high-resolution model system, the precision rapid update forecasting system (PRUFS) model, running for East China at 3 km grid intervals. The variables studied include 2 m temperature, 2 m relative humidity, and 10 m wind speed at 311 standard ground-based weather stations. The dataset for training G2N is a year of historical PRUFS model outputs and the surface observations of the same period and the assessment of the G2N performance are based on the output of two months of real-time G2N. The verification of the real-time results shows that G2N reduced RMSEs of the 2 m temperature, 2 m relative humidity, and 10 m wind speed forecast errors of the PRUFS model by 19%, 24%, and 42%, respectively. Sensitivity analysis reveals that increasing the number of the target stations for simultaneous correction helps to improve the model performance and reduces the computational cost as well indicating that enhancing the loss function with spatial regional meteorological structure is helpful. On the other hand, adequately selecting the size of influencing grid areas of the model input is also important for G2N to incorporate enough spatial features of model forecasts but not to include the information from the grids far from the correcting areas. G2N is a highly efficient and effective tool that can be readily implemented for real-time regional NWP models.

Keywords: deep learning; NWP; post-processing; grid to stations; forecast error correction

Citation: Qin, Y.; Liu, Y.; Jiang, X.; Yang, L.; Xu, H.; Shi, Y.; Huo, Z. Grid-to-Point Deep-Learning Error Correction for the Surface Weather Forecasts of a Fine-Scale Numerical Weather Prediction System. *Atmosphere* **2023**, *14*, 145. https://doi.org/10.3390/atmos14010145

Academic Editors: Xiaolei Zou and Zhemin Tan

Received: 8 December 2022
Revised: 28 December 2022
Accepted: 30 December 2022
Published: 9 January 2023

1. Introduction

Accurate weather forecasting is crucial for the development of society and economy, and human activities and safety. With the rapid development of atmospheric modeling, observation systems, and high-performance computing, numerical weather forecasting capability and accuracy have been improved significantly [1]. Nevertheless, due to the chaotic nature of the weather processes and unavoidable uncertainties in various numerical model components, numerical weather forecasts contain significant errors. Therefore, a correction of model forecast errors is necessary to improve the applications of the model outputs. Several statistical post-processing techniques have been developed for model forecast correction. Among them, model output statistics (MOS) [2] and perfect procedures (PP) [3] are widely used in the current numerical model forecast corrections. While the PP approach achieves a correction by establishing a linear statistical relationship between observations and the NWP model analysis, the MOS method pairs observation data with

the output of NWP and then obtains the correction based on linear regression. In addition, the Kalman filter technique has also been applied for bias correction. Unlike MOS, the Kalman filter technique adjusts its filter coefficients in real time [4,5]. An analog ensemble method proposed by [6] showed an improved capability and has been successfully applied for wind and solar energy forecasting [7–9].

In addition to the statistical model output post-processing, several machine learning techniques have been investigated and have demonstrated benefits and great potential [10]. Li et al. [11] proposed a model output machine learning scheme (MOML) that uses multiple linear regression as well as random forest methods to correct the 2 m temperature in the ECMWF model for the Beijing area. Compared with MOS, which works for single-station site correction, MOML incorporates the spatial and temporal structure of the grid data. Cho et al. [12] used machine learning methods including random forests, support vector regression, artificial neural networks, and multimodel ensembles to establish statistical relationships between predictors and predictands to correct the model forecasts of extreme temperatures in urban areas, demonstrating some ability of the machine learning algorithms for modeling nonlinearities of the weather processes.

In the last decade, convolution neural network (CNN)-based deep-learning technology has made significant strides and offers a natural upgrade to the traditional model output post-processing methods. Rasp et al. [13] proposed a neural network-based model for correcting the 2 m temperature model forecasts in Germany. Han et al. [14] proposed a CU-net model to correct the gridded forecasts of four weather variables of the European Centre for Medium-Range Weather Forecast Integrated Forecasting System global model (ECMWF-IFS): 2 m temperature, 2 m relative humidity, 10 m wind speed, and 10 m wind direction. Their approach turned post-processing into an image transformation problem in the context of image processing. Zhang et al. [15] constructed Dense-CUnet and Fuse-CUnet models based on the CU-net model proposed by Han et al. [14]. By introducing a dense convolution module and a variety of meteorological elements and terrain features into the model, they were able to improve the results of Han et al. [14].

In the last decade, fine-grid numerical weather forecasts with grid intervals of 1–3 km became popular. It is very desirable to explore the deep-learning approaches to extract the meso- and small-scale features of weather circulations simulated by high-resolution numerical models and applied them for model forecast error correction. In theory, multi-scale features of weather circulations contain more information about the model forecast errors that the traditional error correction models, which were based on single-point time-sequence data, could not include. In this study, we developed a grid-to-multipoint (G2N) deep-learning model for correcting the 2 m temperature, 2 m relative humidity, and 10 m wind speed forecasts of a rapid-updating high-resolution weather model (named PRUFS: Precision Rapid Updated Forecast System) at multiple weather stations in East China. Sensitivity tests were conducted to study the impact of the scales of the input area (the model grids) and the target area, i.e., the number of target stations. The former helps to determine the scale of the mesoscale circulation features for the optimal error corrections for the target stations and the latter assesses the effect of the number of target stations for simultaneous error correction with multitasking learning.

2. Data and Method

2.1. Data Description

PRUFS is a precision rapid update forecasting system based on the U.S. Weather Research and Forecasting Model (WRF) and four-dimensional data assimilation (FDDA) technology [16]. The model uses the analysis and forecast fields from the NOAA/NCEP Global Forecast System (GFS) to generate boundary conditions and assimilated various observations in the region during the model initialization. PRUFS runs with a four-nested level with a 3 km grid covering east-central China (Figure 1) since the beginning of 2020. The model system runs at hourly cycles and in each cycle, it generates 24 h forecasts, outputting at hourly intervals.

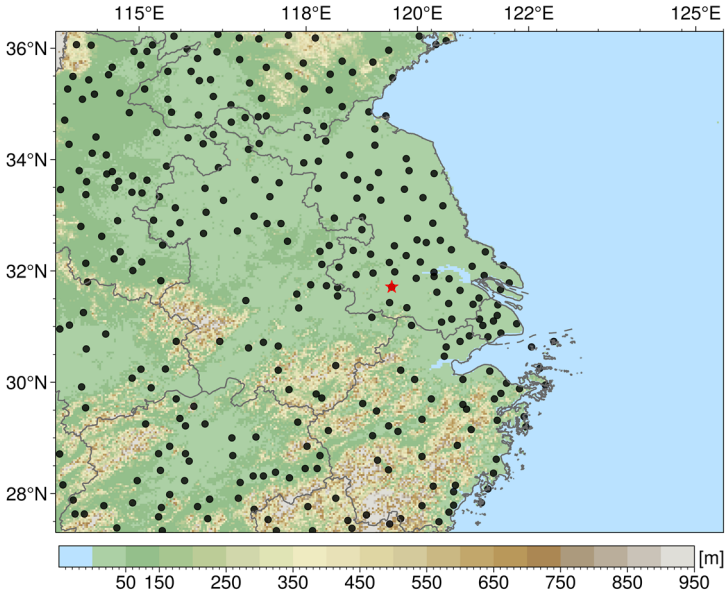

Figure 1. The 3 km horizontal resolution area of PRUFS and the distribution of the 311 automatic weather stations (black dots). The background color shows the height of the terrain, and the red star is the station "Jintan", to be discussed in the later section.

The weather observations were obtained from the state standard ground-based weather stations of the China Meteorological Administration. In this paper, the grid forecast of the PRUFS 3 km domain is corrected by using ground-based meteorological station data. The PRUFS model output and observations are collected for the period from August 2020 to December 2021. The data during the equipment maintenance period September–October 2021 are excluded. The selected computational domain is $(lon \in [113.5^\circ\ \mathrm{E}, 125.5^\circ\ \mathrm{E}],\ \mathrm{lat} \in [27.3^\circ\ \mathrm{N}, 36.3^\circ\ \mathrm{N}])$, with 301×401 grid points. The observation sites and topographic information are shown in Figure 1. The meteorological elements to be corrected are the 24 h hourly forecasts of 2 m temperature (T_2), 2 m relative humidity (RH_2), and 10 m wind speed (W_{10}), generated by PRUFS. With 24 forecast cycles each day, there is a total of 576 model samples (i.e., 24×24) per day.

2.2. G2N Model

Unlike the traditional MOS and PP models that are based on individual station time-sequence weather observations and the model forecast values interpolated at the weather stations (i.e., One-2-One), G2N uses the two-dimensional gridded meteorological information of the model forecasts. Thus, G2N can exploit the two-dimensional gridded meteorological structures (G) of the model forecasts for multiple-site (N sites) weather forecast error correction (i.e., G2N). The gridded model forecast variables can be considered as images in the field of image processing, with the forecast at each grid point corresponding to a pixel. Therefore, the characteristics of meteorological structures can be extracted using image feature engineering technology.

AlexNet is among the simple and effective image-processing deep-learning network models based on a convolutional neural network (CNN). It was first proposed by Krizhevsky et al. [17]. AlexNet consists of five convolutional layers, six pooling layers, and three fully connected layers. The AlexNet model contains a local response normalization layer (LRN) for amplification or suppression of neural activation, and it works together with Dropout methods to prevent the network from overfitting. We construct G2N based on the AlexNet framework.

The change in data distribution due to a change in the network parameters during training is called Internal Covariate Shift [18]. In AlexNet, to avoid the problem of Internal Covariate Shift that slows down convergence and even degrades network performance, a

batch normalization [19] was introduced to replace LRN and Dropout, making the network more robust to the changes in the network parameters and activation functions during the training. BN also solves the problem of gradient disappearance and reduces the negative impact of ICS. The BN layer averages the input values of neurons in the layer to redistribute them to a normal distribution with a mean of 0 and a variance of 1. This allows the increasingly distorted distribution to return to the standard distribution. In this study, to adapt AlexNet for weather variables processing, we replaced LRN and Dropout with BN layers to use it as the core of G2N.

The structure of the G2N model is shown in Figure 2a. G2N takes 2D grid model data as input and has 6 convolutional layers, 4 pooling layers, and 3 fully connected layers. The convolutional layer and pooling layer play a role in extracting 2D features of the model forecasts and the fully connected layer integrates the local information and flattens the feature information. Firstly, G2N convolves and pools the input weather forecast data several times to extract the number of multiscale features. Then, three full-connected operations are performed to finally get the correction results at the weather stations. Each convolutional layer in the figure is followed by a BN layer operation, and then the ReLU activation function is connected. In the fully connected layer, ReLU is also used as the activation function.

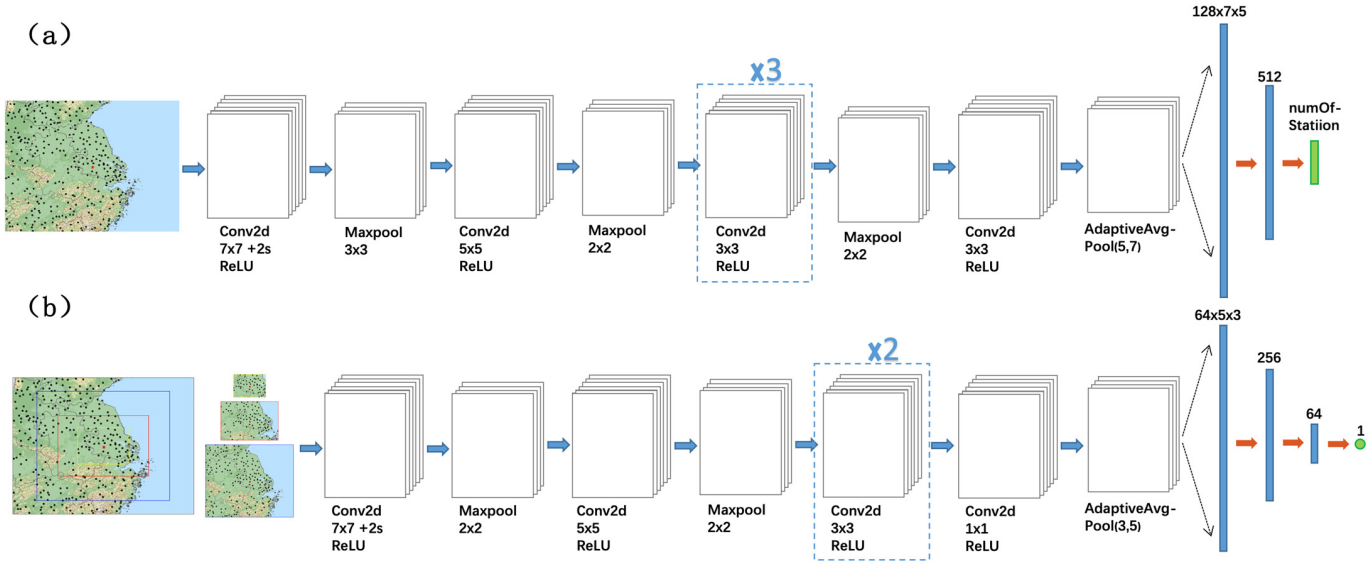

Figure 2. Multisite (G2N, (**a**)) and single-site (G2-One, (**b**)) forecast error correction models for grid forecasting. (**a**) Whole grid area as input for multisite correction of 311 sites. (**b**) Different proportions of grid regions as input for single-point correction experiments.

With the G2N grid-to-station deep-learning architecture, the sizes of model grid forecasts (G) and the numbers of sites (N, weather stations) to be corrected are the two most important model configuration parameters to be considered. Thus, we conducted extensive sensitivity modeling experiments to study the impact of different G and N. Among them, one special case is N = 1, where we let the model work to correct the forecast at only a single site with an input of the model 2D grid data. This is named G2-One (Figure 2b) and is used to study the influence of the G size on the correction results. When changing the size of the input 2D grid data for the G2-One tests, a proper simplification of the G2N model is needed, given in Figure 2b. Adaptive pooling is used in the last layer of the network model, allowing the model to receive flexible sizes of inputs.

In the training process, it is easy to appear over-fitting phenomena with small training errors and large generalization errors. Regularization is a method to solve the overfitting problem in machine learning. The model was trained using L2 regularized weight decay to

control overfitting [20]. Adam was chosen as the optimizer for the gradient descent process of the neural network. The loss function is defined as the mean squared error (*MSE*):

$$Loss = MSE = \frac{1}{M}\frac{1}{N}\sum_{m=1}^{M}\sum_{n=1}^{N}(output_n^m - observe_n^m)^2 \quad (1)$$

where m and n represent the sample number and site number and M and N the batch size and site number, respectively. Each epoch traverses the training set's data multiple times, updating the neural network's parameters with each iteration's batch size. The loss functions of the training and validation datasets are computed after each epoch. In the later examples, all calculations are terminated after 200 epochs of training.

2.3. Data Pre-Processing and Dataset Partitioning

The PRUFS model output and surface observations from August 2020 to August 2021 were selected to construct the training and validation set and the real-time operational data of PRUFS and surface observations from November and December 2021 were used as the test dataset to evaluate the G2N model. The input data of G2N are the PRUFS hourly gridded 0–24 h forecasts, and the training labels are the surface weather observations at the corresponding time of each forecast of a given cycle and forecast time. A bilinear interpolation method was used to interpolate the PRUFS model forecast to 311 surface weather stations to match the observations (Figure 3). The sample data with empty or invalid values were rejected. The interpolated PRUFS forecasts, G2N outputs, and observations are used to calculate the loss function during the G2N model training and evaluate the model performances.

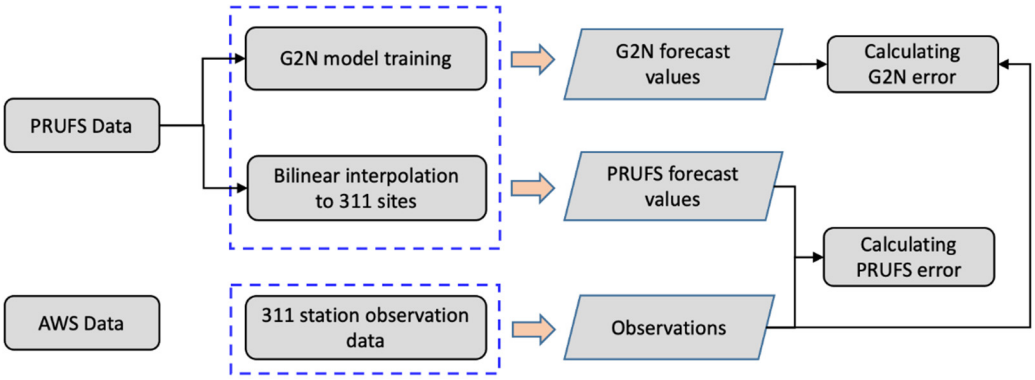

Figure 3. Flow chart of data pre-processing.

After the pre-processing, the number of valid labeled data samples is 167,683,000. The dataset was divided into training and validation sets at an 8:2 ratio. To prevent "information leakage", the 20% validation sets are randomly selected in blocks of continuous 24 forecast cycles, i.e., a whole-day period.

2.4. Model Evaluation Statistics

The model evaluation is conducted by computing the root mean square error (RMSE), systematic bias (BIAS), and the Pearson correlation coefficient (CC), which are the most used metrics for weather forecasting. The root mean square error is calculated as

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{M}\frac{1}{N}\sum_{m=1}^{M}\sum_{n=1}^{N}(output_n^m - observe_n^m)^2} \quad (2)$$

The systematic bias (BIAS) formula is

$$BIAS = \frac{1}{M}\frac{1}{N}\sum_{m=1}^{M}\sum_{n=1}^{N}(output_n^m - observe_n^m) \quad (3)$$

The Pearson correlation coefficient (CC) formula is

$$Corr(output, observe) = \frac{\sum(output - \overline{output})\left(observe - \overline{observe}\right)}{\sqrt{\sum(output - \overline{output})^2\left(observe - \overline{observe}\right)^2}} \quad (4)$$

3. Results and Analysis

3.1. Overall Test Results

Model evaluation is based on the datasets collected during real-time G2N applications along with the PRUFS forecast during November and December 2021. The root mean squared error (RMSE) of 2 m temperature, 2 m relative humidity, and 10 m wind speed at 311 stations of the PRUFS forecast and G2N correction for the 0–24 h forecasts were calculated and the results are presented in Table 1. The percentages of improvement (POI) of the corrected forecast accuracy are also given. It can be found that the convolutional-based G2N model corrected the model forecast errors effectively. The RMSE of all three variables is reduced significantly. The POI of the temperature forecast accuracy increased by 19.4%, the relative humidity by 24.5%, and the wind speed forecast has the greatest enhancement rate of 42.8% by the G2N model.

Table 1. The RMSE and improvement percentages (IP) (see Equation (6)) of the 2 m temperature, 2 m relative humidity, and 10 m wind speed 0–24 h forecasts corrected by G2N, averaged at all stations.

Element	Test Dataset		
	PRUFS	G2N	IP
2 m-T	2.22	1.79	19.4%
2 m-RH	16.25	12.27	24.5%
10 m-WD	1.66	0.95	42.8%

The hour-by-hour errors of the PRUFS forecasts and G2N corrections at each station were computed, and the distribution characteristics of the errors were presented in Figure 4. The error is defined as

$$Error = forecast - observation \quad (5)$$

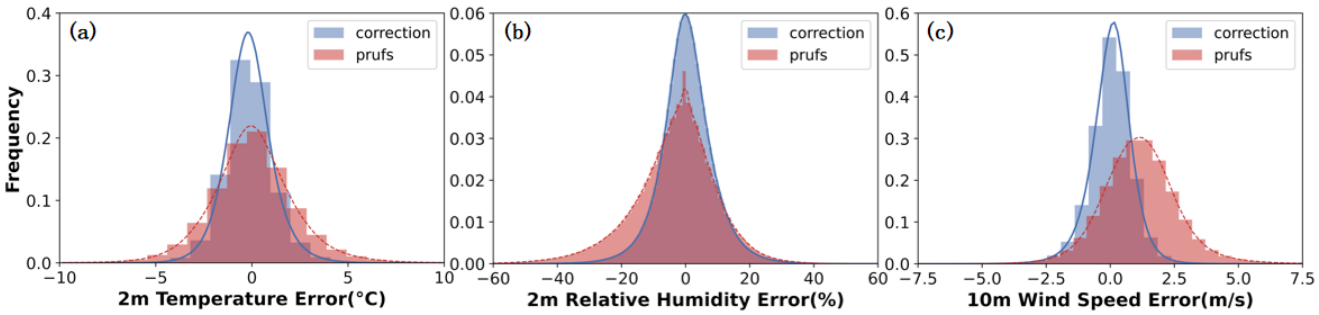

Figure 4. Normalized frequency distributions of the PRUFS forecast errors and the G2N corrected forecast errors. The (**a**) 2 m temperature, (**b**) 2 m relative humidity, and (**c**) 10 m wind speed, with error bins at 1 °C, 1%, and 0.5 m/s, respectively.

Figure 4 shows that the bias of the 2 m temperature, 2 m relative humidity, and 10 m wind speed of the PRUFS forecast are 0.41, −3.43, and 1.15, respectively. After the G2N

correction, they are reduced to −0.15, 0.37, and 0.03, respectively. The distribution of the errors of the G2N-corrected 2 m temperature, 2 m relative humidity, and 10 m wind speed are approximately symmetric about and shrunk to the 0-error point, indicating that G2N is effective in eliminating both negative and positive systematic errors. Notably, the distribution of the 10 m wind speed forecast errors by PRUFS shows an overall apparent positive bias. Several previous studies reported similar results [21–23]. G2N effectively corrected such wind speed biases. The overall systematic errors of the 2 m temperature and 2 m humidity forecasts of PRUFS were not as substantial as the wind. Following the G2N correction, the number of samples with larger temperature and humidity errors also significantly decreased.

To further examine the details of the error properties, density scatter plots of the forecast–observation pairs of the PRUFS forecast and the G2N correction are plotted and the results are shown in Figure 5. The samples include all stations and forecast times during the test period.

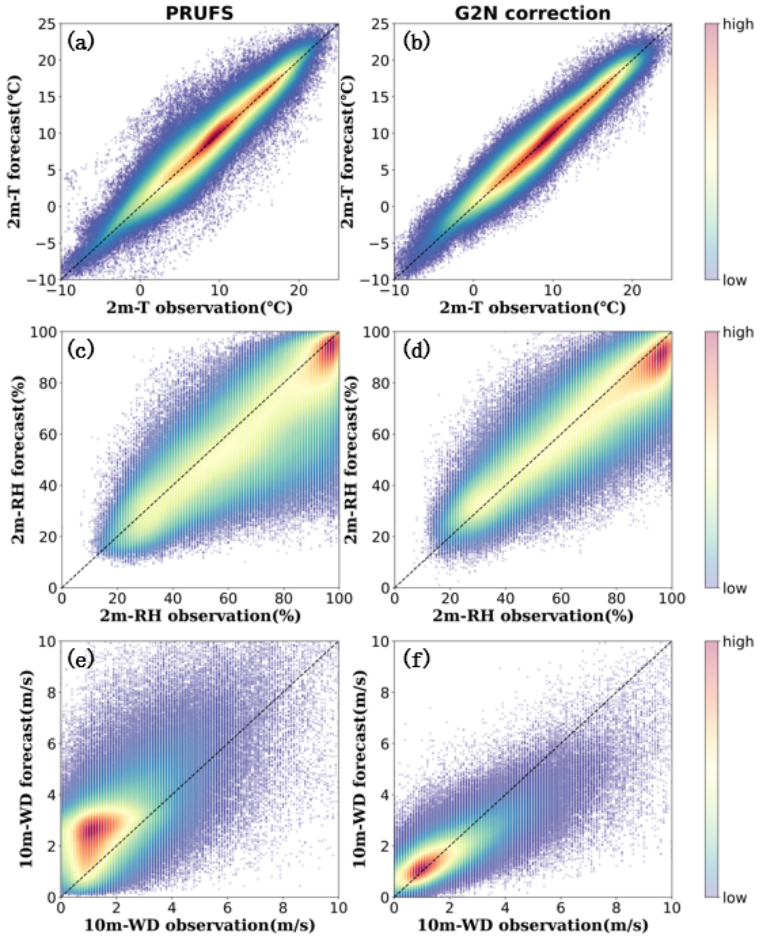

Figure 5. Density scatter plots of the forecast–observation pairs of the PRUFS forecasts (**1st column**) and the G2N correction (**2nd column**). The (**a**,**b**) 2 m temperature, (**c**,**d**) 2 m relative humidity, and (**e**,**f**) 10 m wind speed.

The forecast–observation pairs of all three variables converge more compactly around the black diagonal lines after the correction, i.e., the corrected forecast is closer to the observed value. For example, the variance of the wind speed (Figure 5e,f) is reduced from 2.21 to 0.7. PRUFS underestimated RH (Figure 5c) with the samples skewed to the right of the diagonal and it is removed in the G2N corrected data (Figure 5d), resulting in more centralized and symmetrical error distributions around the diagonal. Similarly, the

10 m wind forecasts (Figure 5e,f) were overall largely overpredicted by PRUFS and G2N dramatically eliminated this bias and the overall errors too.

3.2. Forecast Lead Time and Daily Variation

To analyze the performances of the G2N model for correcting the forecast at different lead times and different times of the day, the samples in the test dataset were grouped according to the forecast lead time and times of day, respectively. After grouping, the error statistics were analyzed for the times in each group. Figure 6a,c,e show the forecast scores of PRUFS and G2N for the 0–24 h lead times. As the forecast lead time increased from 1 to 24 h, the 2 m temperature RMSE increased from approximately 2 to 2.5 °C, the 2 m relative humidity RMSE increased from approximately 12 to 15%, and the 10 m wind speed RMSE increased from 1.8 to 2.0 m/s. After the G2N correction, the RMSE of these three variables is reduced by approximately 1.3 °C, 8%, and 0.8 m/s. Furthermore, G2N can correct the larger errors at longer lead times more effectively for the 24 h forecasts examined here, with the RMSE of the corrected 2 m temperature forecasts increasing only by 0.3 °C, and the corrected 2 m relative humidity and 10 m wind speed errors nearly unchanged with the lead time. This result shows that the G2N model can automatically adjust the magnitude of the error correction according to the error growth for the forecast lead times examined herein.

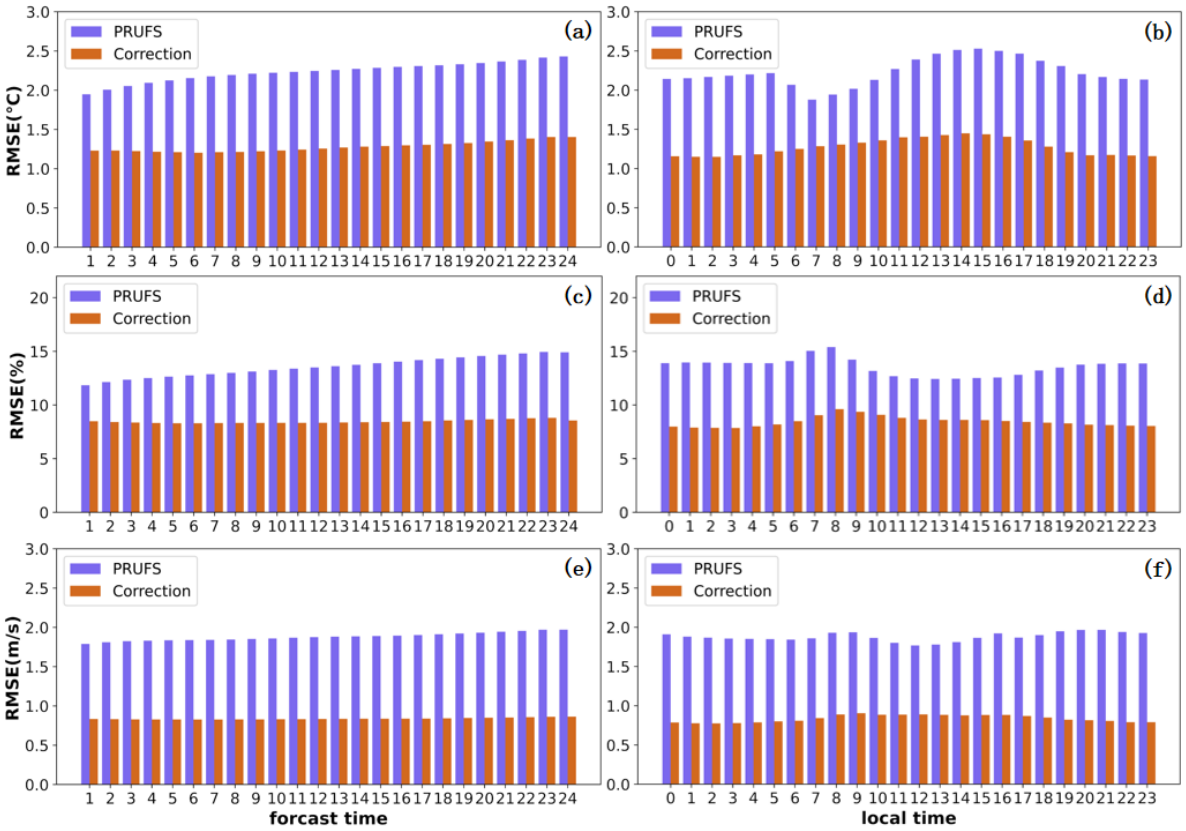

Figure 6. The variation of the RMSE of the PRUFS 0–24 h forecasts and the G2N correction for 2 m temperature (**a**,**b**), 2 m relative humidity (**c**,**d**), and 10 m wind speed (**e**,**f**) with forecast lead time (**left panels**) and diurnal variation (**right panels**); The horizontal coordinate of the left panels is the forecast lead time and that of the right panels is the local time.

The results of the RMSE of the three meteorological variables at different times of day for the PRUFS forecast and G2N correction are shown in Figure 6b,d,f. The errors of the PRUFS forecasts display significant diurnal variations. The RMSE of the 2 m temperature forecasts reached a peak at 15:00 LT, and a minimum at 7:00 LT. The evolution trend of the 2 m relative humidity errors is approximately opposite to the temperature errors, with an

error peak at 8:00 LT, and a valley at around noon LT. The error of 10 m wind speed is less fluctuated. After the G2N corrections, the RMSE of the forecasts of all three variables were significantly reduced at all times of day, with a diurnal variation trend generally consistent with the PRUFS forecasts. This suggests that the diurnal variations of physical processes that caused the PRUFS model errors may also lead to some difficulties for the G2N model.

3.3. Spatial Distribution of the G2N Performances

To analyze the horizontal distribution of the G2N performances, the RMSE of 2 m temperature, 2 m relative humidity, and 10 m wind speed were calculated for each station for all samples of the test dataset. The PRUFS forecast RMSE for all three variables was significantly reduced (Figure 7) at all stations by G2N. The PRUFS temperature prediction errors at several stations were higher than 7.55 °C. After the G2N correction, they were reduced to less than 2.0 °C. In general, the central regions of the domain achieved the best correction results, where the overall error of the PRUFS relative humidity forecast is reduced from ~16% to less than 13% by the G2N model, and the wind speed error from ~1.3–1.5 m/s to below approximately 0.5 m/s. Furthermore, the G2N model is more effective for the stations where the PRUFS forecast errors are larger. At most stations, the G2N model gains IP values over 60% for wind speeds. For relative humidity, there are approximately half of the stations yield 60% IP.G2N performed slightly worse at the southern part of the domain and the northwest corner because the peripheral spatial information of the sites at and across the boundary is not included. The correction effect of wind speed is most effective throughout the domain.

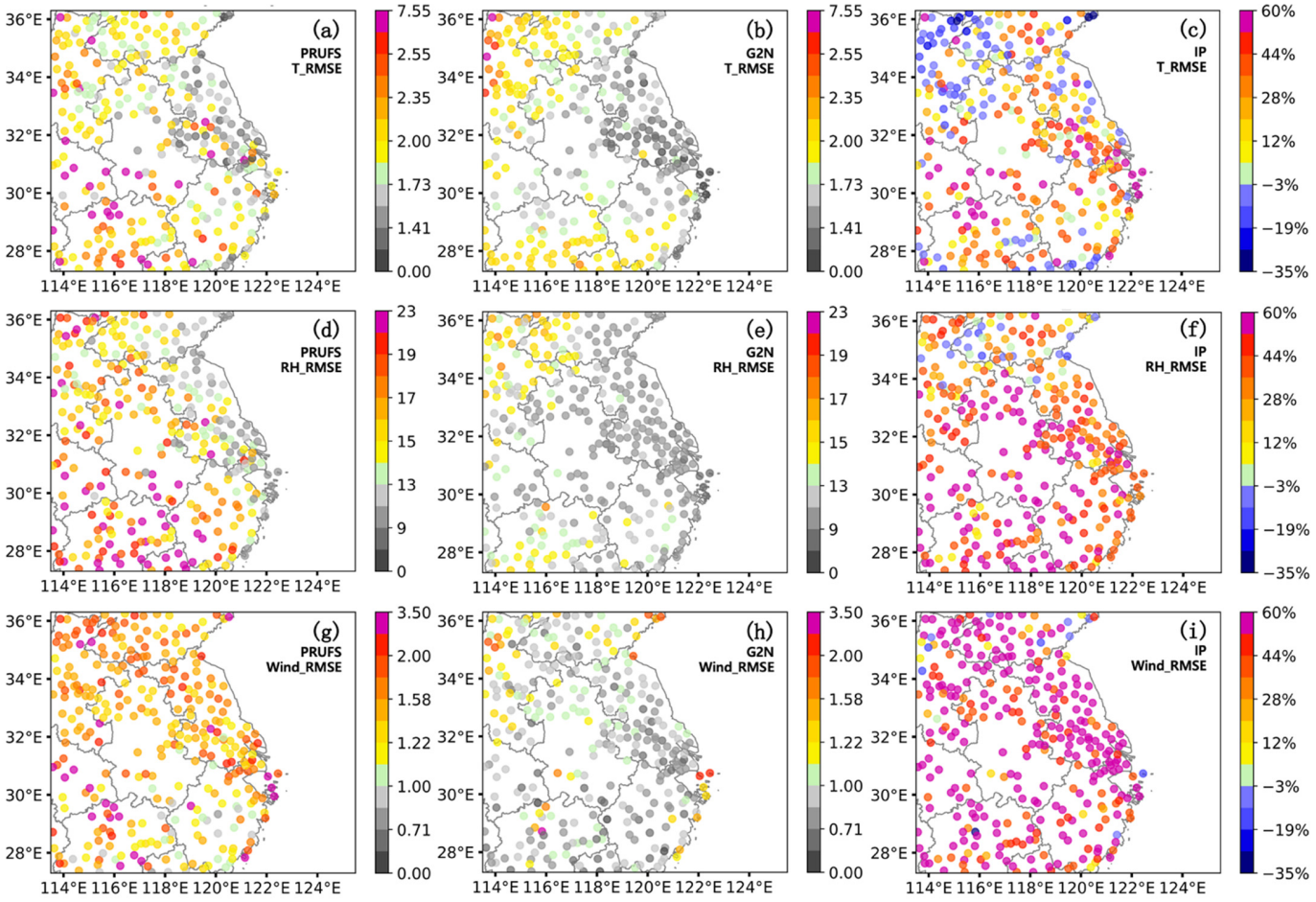

Figure 7. The RMSE of the PRUFS forecasts (**left panels**) and the G2N correction (**middle panels**) and the corresponding G2N improvement percentages (**right panels**, %) of 2 m temperature (**a**–**c**), 2 m relative humidity (**d**–**f**), and 10 m wind speed (**g**–**i**).

To quantitatively compare the G2N effect among the stations, the improvement percentage (IP) of the RMSE of the G2N correction over the PRUFS forecasts was calculated for each site as follows

$$IP = \frac{PRUFS\_forecast_{RMSE} - G2N\_correction_{RMSE}}{PRUFS\_forecast_{RMSE}} \times 100 \quad (6)$$

Figure 7 shows that more than half of the stations gain an IP over 30% for all three meteorological elements although some stations in the northwest marginal area and the southern boundary show a negative effect. A lack of spatial feature information at the edges may impose an unfavorable effect on these sites. Again, the G2N model is most effective for correcting the wind forecast errors, with IPs at most stations larger than 15% and more than a half gained over 50%.

Figures 8 and 9 show the bias and the Pearson correlation coefficients for the forecasts of the three meteorological variables, respectively. The bias of the PRUFS model forecast is significantly reduced by the G2N correction. The PRUFS model temperature forecasts have over 1.3 °C bias at several clustered surface weather stations. They are significantly reduced by the G2N corrections, to less than 0.5 °C. The PRUFS 2 m relative humidity forecast has an overall negative bias (approximately −6.92%) and its wind forecast has a positive bias (approximately 0.95 m/s), and they are reduced to −2.91% and −0.18 m/s, respectively, after the G2N correction.

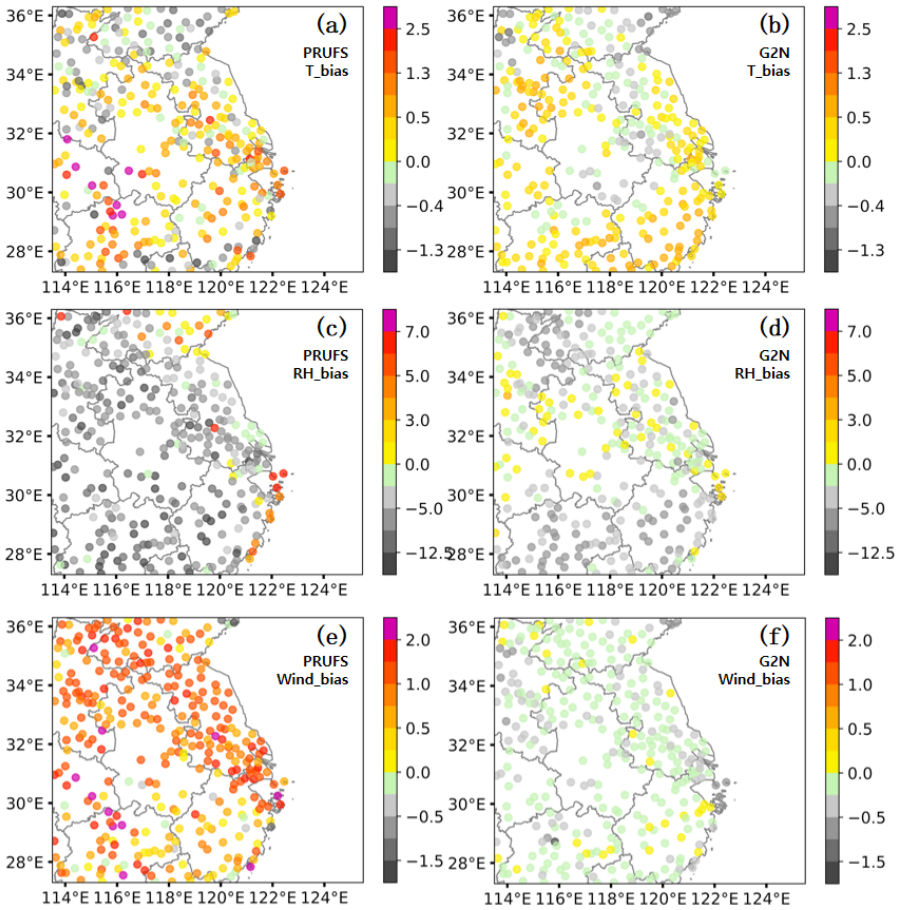

Figure 8. Horizontal distribution of the bias of the PRUFS forecasts (**left panels**) and the G2N correction (**right panels**) of 2 m temperature (**a**,**b**), 2 m relative humidity (**c**,**d**), and 10 m wind speed (**e**,**f**).

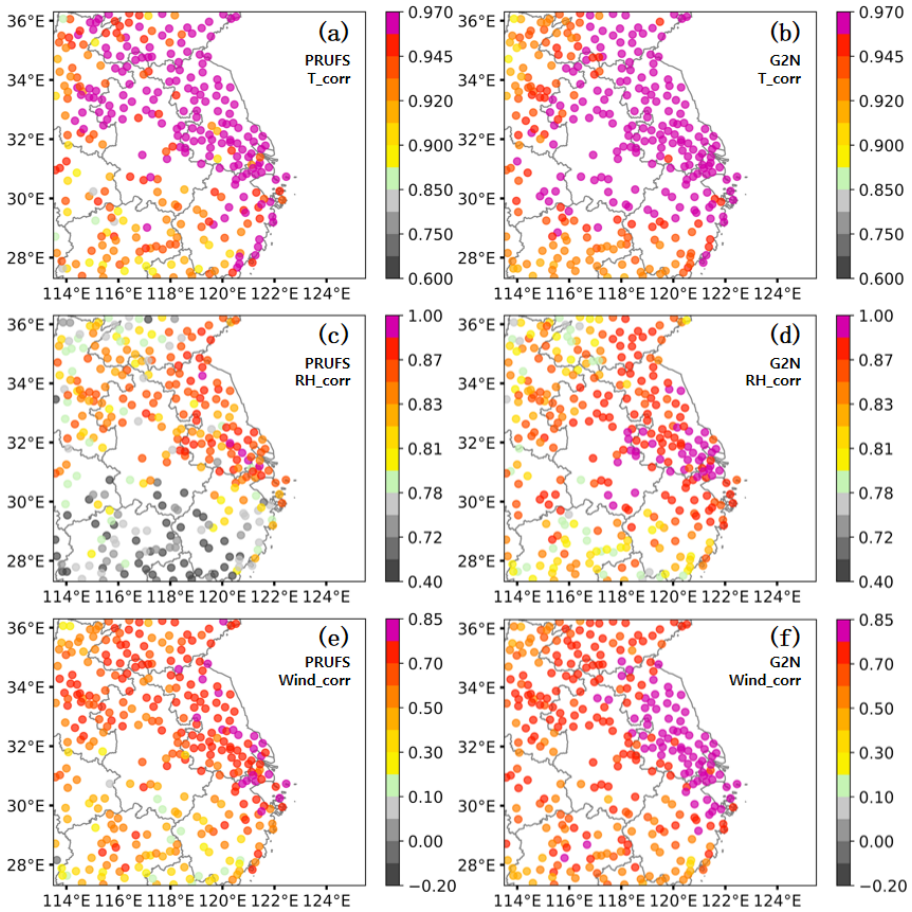

Figure 9. Horizontal distribution of the Pearson correlation coefficients concerning the observations of the PRUFS forecasts (**left panels**) and the G2N correction (**right panels**) of 2 m temperature (**a**,**b**), 2 m relative humidity (**c**,**d**), and 10 m wind speed (**e**,**f**).

In comparison with the PRUFS forecast, the correlation between the G2N corrected forecasts and the observations is also significantly improved for all three meteorological variables (Figure 9). The correlation coefficient (r) can be assessed by the general guidelines proposed by Cohen et al. [24,25], $|r| < 0.3$ is defined as weakly correlated; $0.3 < |r| < 0.6$ as moderately correlated; $0.6 < |r| < 0.8$ as strongly correlated; and $0.8 < |r| < 1$ as extremely strongly correlated.

All station average correlation coefficient for the PRUFS temperature forecast was approximately 0.946 and it reached 0.952 after the G2N correction. For relative humidity, the all-station average correlation coefficient was 0.793 for the PRUFS forecast and ~95% of the stations are strongly correlated. After the G2N correction, the all-station average correlation coefficient was improved to 0.852 and the stations with strong correlation increased to approximately 100%. For the wind, all station average correlation coefficient for the PRUFS forecast was 0.626, the proportion of strong correlation sites is 69%, and the proportion of strong correlation sites was 6.8%. After the G2N correction, all station average correlation coefficient of the corrected sites increased to 0.739, the percentage of strongly correlated sites rose to 92.6%, and the percentage of very strongly correlated sites rose to 33%.

4. Sensitivity Analysis of G2N to the Inputs and Learning Areas

G2N realized the forecast error correction by projecting the PRUFS model grid forecasts to the observation sites. Two natural questions are: what is the optimal size (area) of the gridded input data (G) and what is the proper number of stations for the objective function (loss function)? The size of the input data (G) determines the features of the multiscale characteristics of the PRUFS model forecast that are extracted to infer the information

related to the target site. On the other hand, the number of sites (N) of the objective function is a multitask learning problem [26–29], that is, how many adjacent station sites are optimal for simultaneous learning. This section analyzes these two issues by conducting two groups of sensitivity experiments, briefly, G-exp and N-exp.

4.1. Impact of the PRUFS Forecast Input (G-Exp)

A group of G-exps was conducted to investigate the impact of the PRUFS forecast patch sizes, i.e., the areas of G, as the input of G2N, on the G2N correction. For simplicity, the central station "Jintan" (see Figure 1) was selected as a single site for the correction tests, i.e., G2N with N = 1, briefly, G2-One. The structure of G2-One is shown in Figure 2b.

To keep this paper concise, only the 10 m wind speed correction was presented because the results for the other two variables are similar. The experiments were designed by cropping the domain of the input fields with varying area ratios relative to the large default area chosen and discussed in the previous sections (with a dimension of 301 × 401 grid points). "Jintan" was kept approximately at the centers for all the cropping domains. The area ratio is defined as follows.

$$AreaRatio = Cropped\_dimension/Default\_dimension \quad (7)$$

The G2-One model was trained to correct the 10 m wind speed at the "Jintan" station with AreaRatio = 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 1.0, respectively. (Figure 2b shows the cases of AreaRatios = 0.3, 0.5, 0.7). The RMSE and improvement percentage of the 10 m wind speed at the station of these eight G2-One experiments corrected on the test dataset were computed and given in Table 2. The table also includes the results of the G2N331 model.

Table 2. RMSE of the 10 m wind speed of the PRUFS forecast and the correction by G2-One at Jintan and Nanjing for different input areas and the corresponding IP. The evaluation was done on all test datasets.

AreaRatios of Input Domain	PRUFS Forecasts (RMSE)-"Jintan"	G2-One Correction (RMSE)-"Jintan"	IP	PRUFS Forecasts (RMSE)-"Nanjing"	G2-One Correction (RMSE)-"Nanjing"	IP
1.0 (G2N331)	1.47	0.89	39.5%	1.36	0.97	28.7%
1.0	1.48	0.98	33.8%	1.37	1.05	23.4%
0.9	1.48	0.96	35.1%	1.37	1.04	24.1%
0.8	1.48	0.96	35.1%	1.37	1.02	25.5%
0.7	1.48	0.90	39.2%	1.37	0.98	28.5%
0.6	1.48	0.97	34.5%	1.37	1.01	26.3%
0.5	1.48	0.97	34.5%	1.37	1.01	26.3%
0.4	1.48	0.98	33.8%	1.37	1.02	25.5%
0.3	1.48	0.98	33.8%	1.37	1.02	25.5%

Table 2 shows again that incorporating information from other surrounding sites in the loss functions improves the error correction at Jintan (i.e., G2N outperforms G2-One). Nevertheless, for clarity and simplicity, G-exps for Jintan only is presented. It can be seen in Table 2 that the G2-One performance is improved as AreaRatios (the sizes of G) increase from 0.3 to 0.7, and thereafter, the performance degrades as the AreaRatios continue to increase. **This indicates that selecting the proper sizes of spatial structures/features is important for G2N.** If it is too small, the model will not be able to take in sufficient information on the spatial features of the PRUFS forecasts. On the other side, if the input domain size is too large, it may introduce unnecessary noises and/or information burdens that hinder the G2N training.

In addition to the Jintan station, we also computed the training at other stations located in the central regions of the domain. The RMSE of the wind speed at Nanjing (Table 3)

is smaller than those at Jintan, but the trend of the sensitivity test results with different AreaRatios is consistent with that at Jintan. The results for other stations are similar, but not shown for brevity.

Table 3. Improvement percentages of RMSE for the N-Exps with the G2N model.

(a) 2 m Temperature Statistics Results				
N-exps \ **Verification**	**51**	**101**	**199**	**311**
G2N51	16.4%			
G2N101	14.1%	13.8%		
G2N199	19.8%	19.8%	21.8%	
G2N311	20.8%	20.2%	21.8%	18.9%
(b) 2 m Relative Humidity Statistics Results				
N-exps \ **Verification**	**51**	**101**	**199**	**311**
G2N51	23.7%			
G2N101	24%	25.5%		
G2N199	25.7%	27.5%	24.6%	
G2N311	27.8%	28.9%	25.7%	24.5%
(c) 10 m Wind Speed Statistics Results				
N-exps \ **Verification**	**51**	**101**	**199**	**311**
G2N51	44.5%			
G2N101	46.5%	43.4%		
G2N199	44.5%	41.6%	40.9%	
G2N311	47.4%	44%	43.9%	42.8%

For a possible physical explanation of the optimum size for the G2N model training, we think the mesoscale circulation features are critical. For a given station, the model errors are affected by the mesoscale system over the station and the most important structural features of this mesoscale system should be included for the G2N model input. Thus, the "optimum size" should depend on the size of these most important structural features, which is a few hundred kilometers.

4.2. Impact of the Sites for Multitask Learning(N-Exp)

A set of N-exps is carried out to study the impact of assigning different numbers of surface stations for simultaneous learning, i.e., multitask learning. The stations are selected in regions with "Jintan" approximately at the center (see Figure 1), and the experiments take 51, 101, and 199 sites (Figure 10), respectively.

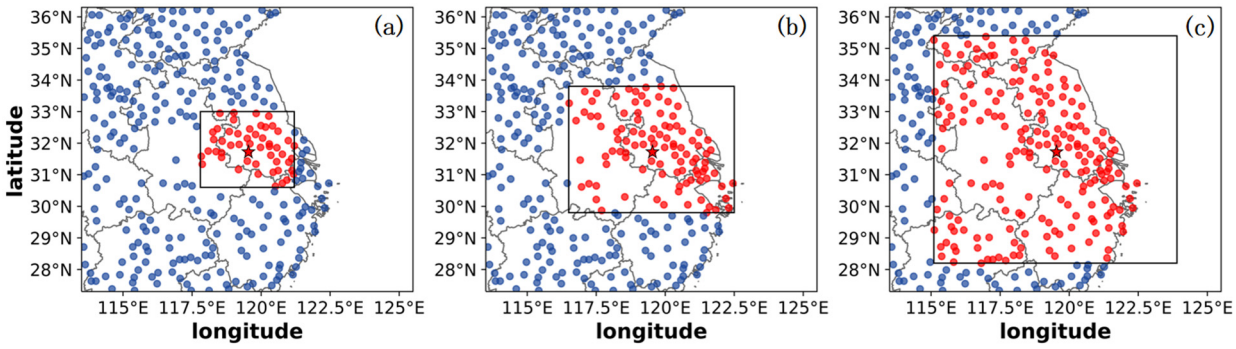

Figure 10. Sub-domains containing 51 (**a**), 101 (**b**), and 199 (**c**) station sites for N-exps. The red star is the station "Jintan".

For N-exp experiments, the G2N model was trained using the same labeled dataset and model forecast input as those discussed in the previous sections, but the loss functions were defined for a varying number of sites (i.e., target domain sizes), i.e., N = 51, 101, 199, and 331, respectively, namely, G2N51, G2N101, G2N199, and G2N311. The performances of these four configurations were assessed based on the outputs of these model runs over the test dataset. The improvement percentages (IP) of RMSE of the G2N outputs, with respect to the PRUFS forecast, were computed for the 51, 101, 199, and 331 site groups, respectively, and presented in Table 3.

The second column of Table 3 labeled as "51" compares the RMSE IP concerning the same 51 sites (shown in red in Figure 10a) corrected by the G2N over the PRUFS forecasts for the four N-exps. The third, fourth, and fifth columns are the same but for the IPs concerning 101, 199, and 311 sites, respectively. It can be seen from Table 3 that for the 51 evaluation sites, as N increases from 51 to 331, the IPs for 2 m T, 2 m RH, and 10 m wind speed gradually grow in general. Similar results can be found for verification statistics computed for 101 and 199 evaluation stations. The learning for all 331 stations achieved the best result. **These results indicate that multistation learning for G2N with more stations is beneficial, not only reducing computing costs dramatically but also increasing the learning skills of the G2N model.**

5. Conclusions

Model output post-processing is a crucial step for correcting the errors of numerical weather prediction. In this study, we established the "grid-to-multipoint" (G2N) convolutional neural network (CNN)-based deep-learning model for correcting the forecast error of an operational high-resolution numerical weather prediction system (PRUFS) running 24 cycles of 0–24 h forecasts, each day over eastern China. G2N corrects model forecast errors by projecting high-resolution weather model gridded forecasts to the surface weather observations. G2N was tested for correcting the forecast of 2 m temperature, 2 m relative humidity, and 10 m wind speed of a high-resolution PRUFS model output. The forecast area contains 311 standard surface weather stations. G2N was trained with one year of data (August 2020 to August 2021) and evaluated by an independent test dataset of the real-time operational PRUFS runs during November and December 2021. The training and testing datasets contain all 24 cycles of 0–24 h forecasts per day. The results show a good performance of G2N for all surface forecast variables corrected and computing efficiency. Furthermore, two groups of sensitivity experiments were conducted to evaluate the impact of changing the input gridded numerical model data sizes and varying the number of stations for multitasking training on the performance of G2N. The main results are as follows.

(1) The G2N model could effectively extract and use the meso- and micro-scale meteorological circulation features, simulated by the high-resolution NWP forecasts, to infer the weather forecast errors at the target stations. The verification of G2N on the test dataset of the 2-month operational runs shows very good improvement percentages of RMSE, 19.0%, 24.5%, and 42.4% for 2 m temperature, 2 m relative humidity, and 10 m wind speed, respectively, in comparison to the PRUFS forecasts.

(2) Sensitivity experiments with selecting mesoscale model forecast (feature) domains show that the size of the input domain has an important impact on the performance of the G2N model. Inputting an excessively small domain will not feed G2N with sufficient spatial features in the PRUFS forecasts that are relevant to the forecast error at the target stations. On the other hand, an excessively large input domain may introduce unnecessary information that hinders the G2N performance.

(3) Sensitivity experiments with multitasking learning strategies (N-exps) show that, for a given input model grid domain, increasing the number of target correction stations within the domain for multitask learning is beneficial to improving the performances of G2N for correcting the errors of all three surface variables. When the three variables (T_2, RH_2, W_{10}) are corrected for the 51 sites, the RMSE improvement percentages of 51 sites

with input threshold are 16.4%, 23.7%, and 44.5%. With the increase in the input threshold, the RMSE improvement percentage of the three variables in 51 sites increased to 20.8%, 27.8%, and 47.4%, with an average increase of approximately 3.8 percentage points.G2N gained the largest error correction when all 311 sites were included in simultaneous learning. This finding indicates that the loss function composed with more target stations could incorporate more relevant spatial loss information and thus increase the G2N model learning abilities.

(4) With its simplicity and high effectiveness, G2N can be readily generalized for post-processing a high-resolution numerical weather prediction system running over other regions. Based on our data and tests, we recommend specifying a patch size of the input model forecast domain with a side dimension of ~600–900 km (200–300 grids) for G2N and including all stations within the domain in the loss function for simultaneous forecast error correction.

The G2N model developed in this paper has been running operationally along with the PRUFS regional numerical weather system to support valuable applications by several customers. For the domain size and stations corrected in this paper, the training time for G2N with one-year samples takes approximately 6 h wall-clock time on a GPU server with Quadro RTX 8000. The G2N real-time run takes only 297 s. Therefore, G2N is a highly efficient and effective tool for post-processing high-resolution NWP forecasts.

Nevertheless, we note that it will be more informative to assess the G2N model performance for a complete year period. Unfortunately, we were not able to access the model data after December 2021. We plan to apply the G2N model for another NWP system in the future and put attention on evaluating the general applicability of the G2N model and its seasonal performance variation characteristics.

We also would like to note that the input for G2N described in this paper only uses a single-element forecast field, e.g., the PRUFS 2 m temperature forecasts for correcting the 2 m temperature at the surface stations. We tested inputting multiple variables, including surface pressure, humidity, and wind, but obtained a degraded performance. Additionally, the results of the present G2N training were obtained without separating the tags of the different forecast lengths, forecast sequences, or forecast cycles of the day. Activating these tags also degraded the G2N performance. Our future work will aim at understanding these limitations and explore more complicated deep-learning models, including refined self-attention algorithms that may amplify the contributions of the key feature in the input and thus gain further improvement on the forecast error correction.

Author Contributions: Conceptualization, Y.L.; methodology, Y.Q., X.J., Y.L. and L.Y.; validation, Y.Q., Y.S. and H.X.; formal analysis, Y.Q., X.J., Y.L. and L.Y.; investigation, Y.Q. and Y.L.; resources, Y.L., Y.S. and H.X.; writing—original draft preparation, Y.Q.; writing—review and editing, Y.L., H.X., Z.H. and L.Y.; visualization, Y.Q., X.J. and Z.H.; funding acquisition, Y.L., H.X. and Y.S.; All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the Science and Technology Grant No.520120210003, Jibei Electric Power Company of the State Grid Corporation of China and partially by the National Key R&D Program of China (Grant No. 2018YFC1507901).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The modeling simulations were carried out on the supercomputer provided by the Nanjing University of Information Science and Technology. The authors thank Peng Zhou for providing and processing the model data analyzed in this study and Xing Wang and Daili Qian for valuable discussions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bauer, P.; Thorpe, A.; Brunet, G. The Quiet Revolution of Numerical Weather Prediction. *Nature* **2015**, *525*, 47–55. [CrossRef] [PubMed]
2. Glahn, H.R.; Lowry, D.A. The Use of Model Output Statistics (MOS) in Objective Weather Forecasting. *J. Appl. Meteorol.* **1972**, *11*, 1203–1211. [CrossRef]
3. Klein, W.H.; Lewis, B.M.; Enger, I. Objective prediction of five-day mean temperatures during winter. *J. Meteorol.* **1959**, *16*, 672–682. [CrossRef]
4. Homleid, M. Diurnal Corrections of Short-Term Surface Temperature Forecasts Using the Kalman Filter. *Weather Forecast.* **1995**, *10*, 689–707. [CrossRef]
5. Delle Monache, L.; Nipen, T.; Liu, Y.; Roux, G.; Stull, R. Kalman Filter and Analog Schemes to Postprocess Numerical Weather Predictions. *Mon. Weather Rev.* **2011**, *139*, 3554–3570. [CrossRef]
6. Delle Monache, L.; Eckel, F.A.; Rife, D.L.; Nagarajan, B.; Searight, K. Probabilistic Weather Prediction with an Analog Ensemble. *Mon. Weather Rev.* **2013**, *141*, 3498–3516. [CrossRef]
7. Alessandrini, S.; Davò, F.; Sperati, S.; Benini, M.; Delle Monache, L. Comparison of the Economic Impact of Different Wind Power Forecast Systems for Producers. *Adv. Sci. Res.* **2014**, *11*, 49–53. [CrossRef]
8. Alessandrini, S.; Delle Monache, L.; Sperati, S.; Cervone, G. An Analog Ensemble for Short-Term Probabilistic Solar Power Forecast. *Appl. Energy* **2015**, *157*, 95–110. [CrossRef]
9. Nagarajan, B.; Delle Monache, L.; Hacker, J.P.; Rife, D.L.; Searight, K.; Knievel, J.C.; Nipen, T.N. An Evaluation of Analog-Based Postprocessing Methods across Several Variables and Forecast Models. *Weather Forecast.* **2015**, *30*, 1623–1643. [CrossRef]
10. Whan, K.; Schmeits, M. Comparing Area Probability Forecasts of (Extreme) Local Precipitation Using Parametric and Machine Learning Statistical Postprocessing Methods. *Mon. Weather Rev.* **2018**, *146*, 3651–3673. [CrossRef]
11. Li, H.; Yu, C.; Xia, J.; Wang, Y.; Zhu, J.; Zhang, P. A Model Output Machine Learning Method for Grid Temperature Forecasts in the Beijing Area. *Adv. Atmos. Sci.* **2019**, *36*, 1156–1170. [CrossRef]
12. Cho, D.; Yoo, C.; Im, J.; Cha, D. Comparative Assessment of Various Machine Learning-Based Bias Correction Methods for Numerical Weather Prediction Model Forecasts of Extreme Air Temperatures in Urban Areas. *Earth Space Sci.* **2020**, *7*, e2019EA000740. [CrossRef]
13. Rasp, S.; Lerch, S. Neural Networks for Post-Processing Ensemble Weather Forecasts. *Mon. Weather Rev.* **2018**, *146*, 3885–3900. [CrossRef]
14. Han, L.; Chen, M.; Chen, K.; Chen, H.; Zhang, Y.; Lu, B.; Song, L.; Qin, R. A Deep Learning Method for Bias Correction of ECMWF 24–240 h Forecasts. *Adv. Atmos. Sci.* **2021**, *38*, 1444–1459. [CrossRef]
15. Zhang, Y.; Chen, M.; Han, L.; Song, L.; Yang, L. Multi-element deep learning fusion correction method for numerical weather prediction. *Acta Meteorol. Sin.* **2022**, *80*, 153–167.
16. Skamarock, W.C.; Klemp, J.B.; Dudhia, J.; Gill, D.O.; Barker, D.M.; Duda, M.G.; Huang, X.-Y.; Wang, W.; Powers, J.G. *A Description of the Advanced Research WRF Version 3*; National Center for Atmospheric Research: Boulder, CO, USA, 2008.
17. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
18. Shimodaira, H. Improving Predictive Inference under Covariate Shift by Weighting the Log-Likelihood Function. *J. Stat. Plan Inference* **2000**, *90*, 227–244. [CrossRef]
19. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *Proc. Mach. Learn. Res.* **2015**, *37*, 448–456.
20. Rahaman, N.; Baratin, A.; Arpit, D.; Draxler, F.; Lin, M.; Hamprecht, F.A.; Bengio, Y.; Courville, A. On the Spectral Bias of Neural Networks. *Proc. Mach. Learn. Res.* **2018**, *97*, 5301–5310.
21. Pan, L.; Liu, Y.; Roux, G.; Cheng, W.; Liu, Y.; Hu, J.; Jin, S.; Feng, S.; Du, J.; Peng, L. Seasonal Variation of the Surface Wind Forecast Performance of the High-Resolution WRF-RTFDDA System over China. *Atmos. Res.* **2021**, *259*, 105673. [CrossRef]
22. Shi, J.; Liu, Y.; Li, Y.; Liu, Y.; Roux, G.; Shi, L.; Fan, X. Wind Speed Forecasts of a Mesoscale Ensemble for Large-Scale Wind Farms in Northern China: Downscaling Effect of Global Model Forecasts. *Energies* **2022**, *15*, 896. [CrossRef]
23. Zeng, X.-M.; Wang, M.; Wang, N.; Yi, X.; Chen, C.; Zhou, Z.; Wang, G.; Zheng, Y. Assessing Simulated Summer 10-m Wind Speed over China: Influencing Processes and Sensitivities to Land Surface Schemes. *Clim. Dyn.* **2018**, *50*, 4189–4209. [CrossRef]
24. Minton, P.D.; Cohen, J. Statistical Power Analysis for the Behavioral Sciences. *J. Am. Stat. Assoc.* **1971**, *66*, 428. [CrossRef]
25. Cohen, J. *Statistical Power Analysis for the Behavioral*; Routledge: London, UK, 1988.
26. Kumar, A.; Daumé, H., III. Learning Task Grouping and Overlap in Multi-Task Learning. *arXiv* **2012**, arXiv:1206.6417.
27. Gao, Y.; Ma, J.; Zhao, M.; Liu, W.; Yuille, A.L. NDDR-CNN: Layerwise Feature Fusing in Multi-Task CNNs by Neural Discriminative Dimensionality Reduction. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 3200–3209.

28. Crawshaw, M. Multi-Task Learning with Deep Neural Networks: A Survey. *arXiv* **2020**, arXiv:2009.09796.
29. Li, Y.; Lang, J.; Ji, L.; Zhong, J.; Wang, Z.; Guo, Y.; He, S. Weather Forecasting Using Ensemble of Spatial-Temporal Attention Network and Multi-Layer Perceptron. *Asia Pac. J. Atmos. Sci.* **2021**, *57*, 533–546. [CrossRef]