atmosphere

MDPI

Article

Improvement of Maximum Air Temperature Forecasts Using a Stacking Ensemble Technique

Linna Zhao [1,2,*], Shu Lu [3] and Dan Qi [4]

[1] State Key Laboratory of Severe Weather and Institute of Artificial Intelligence for Meteorology, Chinese Academy of Meteorological Sciences, Beijing 100081, China
[2] Union Centre for Extreme Weather, Climate and Hydrogeological Hazards, China Meteorological Administration, China University of Geosciences, Wuhan 430074, China
[3] Hunan Meteorological Observatory, Changsha 410118, China; lushu0818@163.com
[4] National Meteorological Center, Beijing 100081, China; qidan@cma.gov.cn
[*] Correspondence: zhaoln@cma.gov.cn

Abstract: Due to the influence of complex factors such as atmospheric dynamic processes, physical processes and local topography and geomorphology, the prediction of near-surface meteorological elements in the numerical weather model often has deviation. The deep learning neural networks are more flexible but with high variance. Here, we proposed a stacking ensemble model named FLT, which consists of a fully connected neural network with embedded layers (ED-FCNN), a long short-term memory (LSTM) network and a temporal convolutional network (TCN) to overcome the high variance of a single neural network and to improve prediction of maximum air temperature. The case study of daily maximum temperature forecast evaluated with observation of almost 2400 weather stations shows substantial improvement over that of single neural network model, ECMWF-IFS and statistical post-processing model. The FLT model can more effectively improve the forecast bias of the ECMWF-IFS model than that of any of the above single neural network model, with the RMSE reduced by 52.36% and the accuracy of temperature forecast increased by 43.12% compared with the ECMWF-IFS model. The average RMSEs of the FLT model decreases by 8.39%, 1.50%, 2.96% and 16.03%, respectively, compared with ED-FCNN, LSTM, TCN and the decaying average method.

Keywords: deep learning; stacking ensemble; post-processing; bias correction

check for updates

Citation: Zhao, L.; Lu, S.; Qi, D. Improvement of Maximum Air Temperature Forecasts Using a Stacking Ensemble Technique. *Atmosphere* **2023**, *14*, 600. https://doi.org/10.3390/atmos14030600

Academic Editor: Graziano Coppa

Received: 22 December 2022
Revised: 13 March 2023
Accepted: 15 March 2023
Published: 21 March 2023

1. Introduction

Surface maximum air temperature plays an important role in people's life, climate prediction, [1] and hydrologic forecasting, etc. [2]. Therefore, reliable forecasts of maximum air temperatures are essential to prevent heat-related disasters, efficiently mitigate the damages caused by high-temperature disasters and appropriately respond to them [3]. Deviations usually exist in the prediction of near-surface elements from numerical models due to complex factors such as atmospheric dynamic processes, physical processes, local topography and geomorphology. In particular, the deviation between the prediction and observation of daily maximum temperature is relatively larger when the weather changes drastically [4]. Therefore, it is still a challenge to realize refined and accurate forecasts for daily maximum temperature. Numerical weather prediction (NWP) models predict surface air temperature by using an atmospheric system containing complex dynamics and land-atmosphere-ocean coupling. In the last 50 years, the NWP has achieved great success with the development of computer technology, modeling techniques and observations [5,6]. However, due to the imperfection of initial values, assimilation methods, systematic errors and physical processes in the model, the uncertainties of the atmosphere cannot be thoroughly described [7,8], resulting in inevitable errors of numerical predictions. Hence, there is relatively limited space to enhance the NWP by improving the physical processes in numerical models or increasing the spatial resolutions. In order to reduce these model biases and further improve the accuracy of numerical model prediction products, various post-processing methods have been developed to correct the errors of NWP models [9,10]. In the early years, the most common post-processing methods included the perfect prediction method (PPM) [11], the model output statistics (MOS) method [12], the Kalman filter (KF) [13], decaying average method, Bayesian model averaging (BMA) [14], frequency matching method and scoring optimization correction method, etc. These methods have been widely used and achieved good results. Specifically, as an objective weather forecast technology, the MOS method has been widely used in the correction of prediction errors for surface wind, precipitation probability and maximum temperature, and formed the operational forecasts [15].

The MOS method establishes a regression model between the observation data and model outputs, which uses the statistical relationship to transform a single-valued model forecast into another single-valued model forecast for more accurate forecasts. However, due to the prediction errors in NWP model outputs [10,12,16], the established equation for error correction based on the statistical relations will also produce errors. Therefore, the effect of bias correction using the MOS method will be limited. As each numerical prediction model is continuously updated, it will invalidate a large number of historical model output data used in the MOS method, leading to the failure of MOS to obtain the latest observation features, which is another limitation of this method. Moreover, some statistical correction methods require certain statistical assumptions. For example, the BMA method assumes prior probability, and different prior probability assumptions of BMA will lead to different or even opposite results. In a word, different statistical models have their own limitations. The available weather forecast is generally limited to within about 10 days due to the chaos in the atmosphere, and the effect of error correction for both traditional and new techniques will be weakened with the extension of leading time.

In recent years, studies have shown that compared with the traditional statistical post-processing technology, the artificial intelligence-based post-processing technology in medium-term numerical prediction model has own advantage in that it is a data-driven method. This technology can implicitly extract the spatio-temporal variations of nonlinear and multi-scale physical relationships from multi-source data, thus significantly enhancing the level of medium- and short-term weather forecasts [17–20]. Machine learning approaches can handle a large number of input variables because they are not sensitive to the multi-collinearity of the input variables [21]. In addition, machine learning can be used to establish a model that works for multiple stations, unlike MOS and KF that require bias correction to build a model for each station. Rasp et al. [17] applied the ANN to the post-processing of ensemble forecasts, and conducted a case study on the 2-m temperature prediction at German surface stations. The results show that the neural network method performs significantly better than the post-processing methods such as the ensemble MOS (EMOS) and quantile regression forests etc., and it is more computationally efficient. In addition, the temporal convolutional network (TCN) is also an emerging network structure suitable for time series [22], and Hewage et al. [23] proposed a lightweight data-driven weather forecasting model by using the TCN. Han et al. [24] transformed the forecast correction problem into an image-to-image translation problem in the field of computer vision. They applied the CU-net (Component Unmixing Network) architecture to the gridded forecasts of 2-m temperature, 2-m relative humidity, 10-m wind speed and 10-m wind direction from the European Centre for Medium-Range Weather Forecasts Integrated Forecasting System (ECMWF-IFS) for error correction, and achieved significant correction performance. The model output machine learning (MOML), a post-processing method for gridded temperature forecasts, matches NWP forecasts against observations through a regression function and shows a better numerical performance than the ECMWF model and MOS, especially in the winter of Beijing [19]. However, a weakness of MOML is the intricate pre-processing associated with feature engineering. Chen et al. [25] proposed an end-to-end post-processing method based on a deep convolutional neural network, which directly learns the mapping relationship between the model predicted field and the observed temperature field and obtains relatively accurate temperature predictions. Recently, Zhao et al. [26] applied a fully connected neural network model with embedded layers for prediction of daily maximum air temperature, reducing the overall root mean square error (RMSE) of the ECMWF-IFS model from 2.746 °C to 1.433 °C.

Due to the complex atmosphere-surface interaction, a single machine learning method is unable to consistently and effectively remove the bias in the NWP model. In fact, the different regressor in machine learning can significantly affect performance of the prediction [27–30]. Single machine learning models, especially the deep learning networks, suffer from the "bias-variance" trade-off [31], and the ensemble machine learning can be an effective way to address this issue. A successful way to reduce the high variance of neural network models is to train multiple models instead of a single model and integrate the predictions of these models, i.e., integration learning. Integration learning not only reduces the variance of predictions, but also yields better predictions than any single model.

Recently, several researchers have attempted to improve the performance of the forecasting by combination (i.e., ensemble) of the diversified machine learning [32–34]. The random forest, support vector regression and ANN, as well as the multiple-model ensemble (MME) were employed to improve the forecast bias of daily maximum temperature in Seoul from the Local Data Assimilation and Prediction System model outputs [21]. Chen et al. [35] applied linear regression models to integrate the ANN and the long short-term memory-full convolutional network (LSTM-FCN) to reduce prediction bias of 2-m temperature at 301 weather stations in mainland China. These experiments have all demonstrated that combining different machine learning models can improve the forecast performance by overcoming the drawbacks of each individual classifier. However, the complexity of the neural networks used in these studies was limited, and the ensemble methods used were relatively simple.

Here, we used the cross-validation method to determine the model hyperparameters, used the mean absolute error (MAE) as the loss function, and used the early stop method to train the model to demonstrate how to use the stacked generalization technology to integrate different types of neural networks. The proposed integrated neural networks are used in the post-processing of daily maximum air temperature forecasting within a neural network framework to reduce the forecast bias of individual neural network models. The daily maximum air temperature forecasting method established in this study will fill the gap that the ECMWF-IFS model output only has 2 m temperature forecast and no daily maximum temperature forecast. Specifically, we explore a case study of daily maximum temperature forecast for 1 January–31 December 2020. We compare the forecast capability between the integrated neural network and the individual neural network, and also compare the forecast capability between the neural network model and the statistical post-processing model. Our ultimate goal is to propose an effective, reliable, and robust ensemble neural network model.

The rest of this paper is organized as follows. The study area and data are briefly described in Section 2. Section 3 introduces the methodology, including the machine learning algorithms, data processing, specified evaluation metrics and experimental schemes. Section 4 presents the results. Finally, the discussions and conclusions are given in Section 5.

2. Data and Feature Selection

2.1. Data

In this work, we focus on the forecasting of daily maximum air temperature with a leading time of 24 h at weather station in Chinese mainland. The study area is Chinese mainland, located within the latitudes of 15°–55° N and the longitudes of 73°–135° E (Figure 1), and there are more than 2400 meteorological stations in the domain.

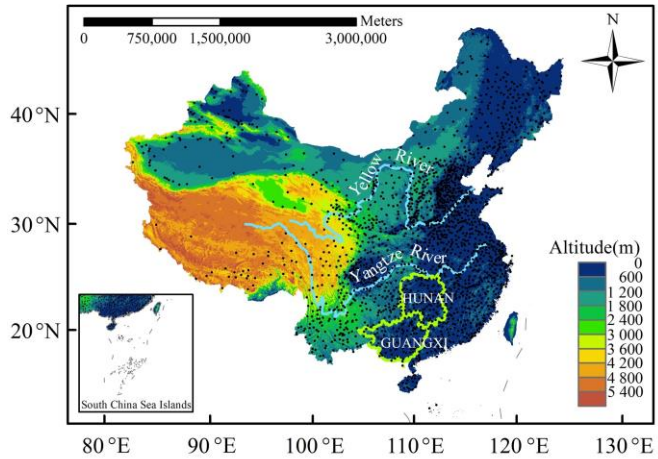

Figure 1. The illustration of the research area and the location of weather stations (black dots).

We used forecast data from the ECMWF-IFS model from 15 January 2015 to 31 December 2020. The forecast data are issued at 1200 UTC, and forecast lead time is from 0 to 72 h with 3-h intervals. The horizontal resolutions of the ground and upper -level air forecasts are 0.125° × 0.125° and 0.25° × 0.25°, respectively. Several predictors (e.g., land-sea mask) have the same value and do not change with time. In addition to these unnecessary variables, 37 primary predictors are chosen, broadly based on meteorological intuition. Table 1 shows these primary predictors and their abbreviations. For comparison with the station observations, the gridded forecasted data are bilinearly interpolated to the weather station locations.

The daily maximum air temperature is available from the National Meteorological Information Center, China Meteorological Administration. Data scrubbing and quality control are performed on the observation and forecast data, and 2238 stations are obtained after excluding a small number of stations containing missing values to ensure the integrity of the observation time series.

Table 1. The primary predictors taken from the ECMWF-IFS model and their abbreviations.

Predictor	Abbreviation	Predictor	Abbreviation
Surface level			
Maximum of temperature *	2Tmax *	Total column water vapor	TCWV
2-m temperature	2T	Large scale precipitation	LSP
2-m dewpoint temperature	2D	Convective precipitation	CP
Skin temperature	SKT	Total cloud cover	TCC
Maximum temperature at 2 m in the last 6 h	MX2T6	Low cloud cover	LCC
Minimum temperature at 2 m in the last 6 h	MN2T6	Forecast albedo	FAL
Sea surface temperature	SST	Snow density	RSN
10-m zonal wind component	10U	Snowfall	SF
10-m meridional wind component	10V	Snow depth	SD
100-m zonal wind component	100U	Convective effective potential energy	CAPE
100-m meridional wind component	100V	Mean sea level pressure	MSL
Total precipitation	TP	Level height of zero-degree layer	DEG0L
Total column water	TCW	-	-

Table 1. *Cont.*

Predictor	Abbreviation	Predictor	Abbreviation
500 hPa, 700 hPa and 850 hPa high level			
Temperature	T	Divergence	D
Zonal wind component	U	Meridional wind component	V
Geopotential height	HGT	Specific humidity	Q
Potential vorticity	PV	Relative humidity	R
Vertical velocity	W	-	-
Calculated predictor			
Vorticity at 500 hPa	VOR500	Vorticity advection at 500 hPa	VOR_ADV500
Vorticity at 700 hPa	VOR700	The temperature advection difference between 850 hPa and 500 hPa	T_ADV850_500
Vorticity at 850 hPa	VOR850	-	-

* The 2Tmax is the maximum forecast of the 2 m maximum temperature in the past 6 h at lead time of 6 h, 12 h, 18 h and 24 h, respectively.

2.2. Feature Selection

The principle of feature selection is that selected feathers should contain as much available information as possible and without making the model complex. The feature selection method used in this study is a combination of variance filtering method, correlation analysis method and mutual information value method. On the basis of primary predictors (Table 1), firstly, the low-variance features are eliminated through the variance filtering method, and the threshold level is set as 0.9 to remove 90% features with the same elements. After filtering the low-variance feature, on the one hand, the factors with the correlation coefficient greater than 0.3 and passing the significance test at the significance level of 0.05 are selected through the correlation analysis method. In this way, the features with a certain linear relationship with the target can be selected. On the other hand, the mutual information method is used to select the features that have a certain nonlinear relationship with the target. The mutual information value between the target and the feature is calculated. The mutual information value can measure the degree of interdependence between two random variables. When the information of a random variable is obtained, the degree of uncertainty reduction in another variable can be determined by the mutual information value. Then, the features of the top 15 predictors are retained with mutual information value from high to low. Finally, the features selected by the correlation analysis method and the mutual information value method together constitute the features which input to the neural network model.

The forecast variables of the ECMWF-IFS model after feature selection, such as the daily maximum temperature, 2-m temperature, etc., with 19 in total. In addition to the ECMWF-IFS model predictors, auxiliary variables are also retrieved to make the model better mine the temporal and spatial information in the predictors. These were chosen broadly based on meteorological intuition. Table 2 shows these feathers.

In this study, the forecast and observation data are divided into the training set, validation set and test set, covering the periods from 15 January 2015 to 30 September 2019, from 1 October to 31 December of 2019 and from 1 January to 31 December of 2020, respectively.

Table 2. The descriptions and abbreviations of all features.

No.	Feather	Abbreviation
1	Maximum of temperature	2Tmax
2	2-m temperature	2T
3	2-m dewpoint temperature	2D
4	Skin temperature	SKT
5	Maximum temperature at 2 m in the last 6 h	MX2T6
6	Minimum temperature at 2 m in the last 6 h	MN2T6
7	Total column water	TCW
8	Total column water vapor	TCWV
9	Convective effective potential energy	CAPE
10	Mean sea level pressure	MSL
11	Level height of zero-degree layer	DEG0L
12	Temperature at 850 hPa	T850
13	Temperature at 700 hPa	T700
14	Temperature at 500 hPa	T500
15	Specific humidity at 850 hPa	Q850
16	Specific humidity at 700 hPa	Q700
17	Specific humidity at 500 hPa	Q500
18	Geopotential height at 700 hPa	Q700
19	Geopotential height at 500 hPa	Q500
20	Daily maximum temperatures lagged the forecast target date by 1 day	2Tmax_lag1
21	Daily maximum temperatures lagged the forecast target date by 2 days	2Tmax_lag2
22	Longitude station	Station_lon
23	Latitude of station	Station_lat
24	Altitude of station	Station_alt
25	Identification serial number of station	Station_ID
26	Season	Season
27	Month	Mon

3. Methods

3.1. Construction of Ensemble Neural Network Model

Due to the randomness of initial weight parameters, neural network models generally have high variances. Using ensemble models can effectively reduce the variance and enhance the model generalization ability [36]. Stacked generalization method is one of the effective methods to construct ensemble models. By integrating multiple models, a better generalization performance can be obtained compared with a single model. Deep learning is a powerful tool to deal with complex nonlinear problems and has started to play a role in weather forecasting. However, the accuracy and dependence of neural network model depend on the different types of neural network adopted, the input features, the methods of feature processing, the addition of auxiliary variables and the hyperparameter settings of the model, which are often required to be selected and designed by the model designer according to the object of modeling prediction. In this study, an ensemble neural network model is constructed based on three neural network models of the ED-FCNN, the LSTM and the TCN by using the stacking ensemble technique (hereafter referred to as the FLT). Usually, the stacking method is divided into the following two levels. Leve-0 trains several different machine learning models to obtain the predictions of each model for the target separately. Leve-1 takes the output result of each model trained in level-0 as the input element and then trains a new model as a level-1, and the output of the level-1 is taken as the final forecast.

In order to reduce the variance of a single neural network model, in this study, as shown in the Figure 2, the ensemble learning FLT model is proposed, which consisted of ED-FCNN, LSTM and TCN by using a stacking approach. The first level of ensemble learning takes the ED-FCNN, LSTM and TCN as the Level-0 models, and the second level uses the fully connected neural network (FCNN) as the Level-1. The FCNN is selected in this paper as the Level-1 model because the FCNN has the characteristics of multiple inputs, so it can process multiple neural networks in the Level-0 sub-network, and then learn how to best integrate the predictions from each sub-neural network model, and it allows the stacked ensemble model to be regarded as a single large model. Figure 3 is the flow chart of this study.

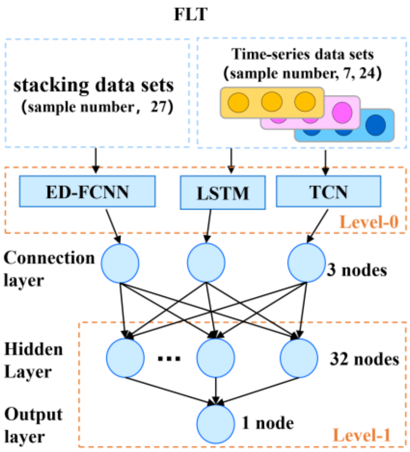

Figure 2. Architectures of the ensemble neural network model (FLT) constructed in this study.

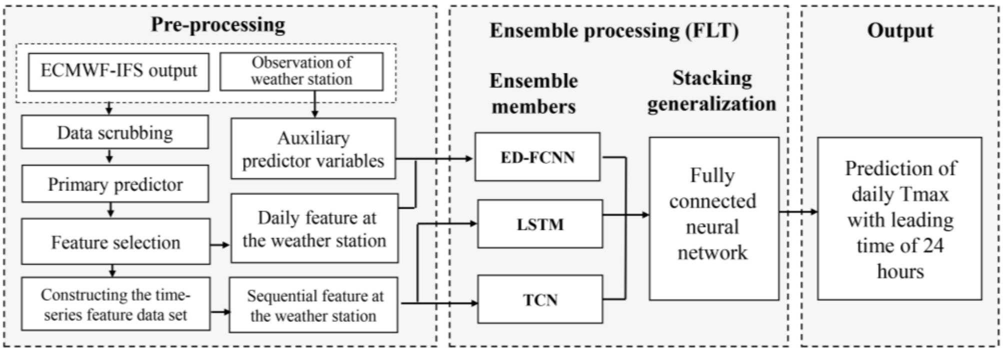
Pre-processing
ECMWF-IFS output
Observation of weather station
Data scrubbing
Primary predictor
Feature selection
Constructing the time-series feature data set
Auxiliary predictor variables
Daily feature at the weather station
Sequential feature at the weather station
Ensemble processing (FLT)
Ensemble members
Stacking generalization
ED-FCNN
LSTM
TCN
Fully connected neural network
Output
Prediction of daily Tmax with leading time of 24 hours

Figure 3. Flow chart of the proposed FLT base on three neural network models by stacking generalization technique.

One of the sub-neural network models at level-0 of the integrated neural network FLT model in this study is ED-FCNN, which is a fully connected neural network with embedding layer. Spatial and temporal information in the features need to be input into the integrated model FLT. In order to process multi-category variables and integrate the spatio-temporal information into the neural network, a fully connected neural network with an embedding layer is induced in this study, namely, ED-FCNN [26], which can be adapted to general regression tasks and does not require the data to have a temporal sequence relationship. For the ED-FCNN, the input features are numbered 1 to 27 in Table 2, which are forecast factors of the ECMWF-IFS model after feature selection, the auxiliary variables and time lagged variables. The input data of all stations is stacked in the ED-FCNN. As these data are in obvious chronological order, the flags of the season and month to identify the temporal information in the data sample are added when constructing the input feature data set for the ED-FCNN. There are two input layers for the ED-FCNN. The first input layer transmits three categorical variables of station number, month and season, where the station number and month are transmitted into the embedded layer after the process of label encoding, and the variable of season is processed by the one-hot encoding. The second input layer passes various meteorological factors and other features, and finally the connection layer connects all the above features.

The other two sub-neural network models at level-0 of the integrated neural network FLT model are Long Short-Term Memory Network (LSTM) (Figure 4a) and Temporal Convolutional Network (TCN) (Figure 4b), respectively. The LSTM is a kind of RNNs, which is specifically used to deal with time series. Compared with the ordinary structure of RNN, LSTM can solve problems such as the long-term dependence of sequence and gradient vanishing. LSTM is also composed of the input layer, hidden layer and output layer. A unique gating mechanism is added to the LSTM cell in the hidden layer, and the information will be updated when the information passes through the LSTM cell. The TCN, is a neural network architecture proposed by Bai in 2017 [22]. It migrates the convolutional neural network (CNN) to time-series applications and mainly applies one-dimensional convolution. Compared with the LSTM and other RNNs neural networks, TCN has the advantage of large-scale parallel processing. It further integrates the technologies of causal convolutions, dilated convolutions and residual connections on the basis of one-dimensional convolution [22]. The causal convolution mainly ensures that the future information will not be leaked during the prediction, which is in line with the characteristics of time-series prediction.

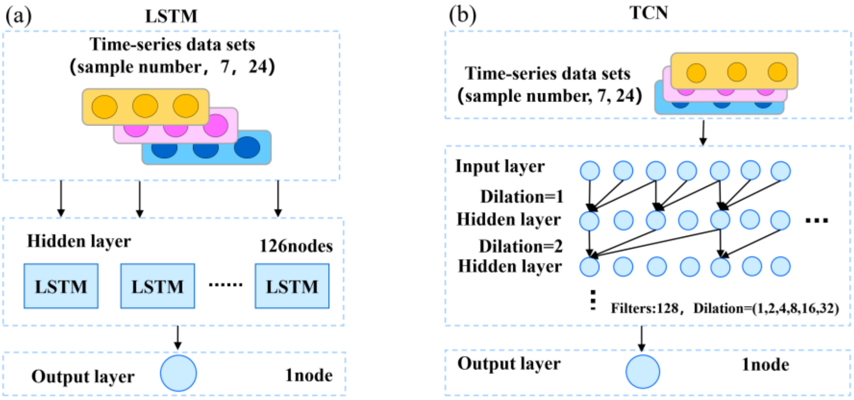

Figure 4. Architectures of the sub-neural network model in an ensemble neural network (**a**) long short-term memory (LSTM) and (**b**) temporal convolutional network (TCN).

For LSTM and TCN constructed in this study, the time-series data set with a fixed form should be constructed as the model input, and the shape of the input array includes the batch size, time-series length and feature number. The steps of constructing the time-series data set are as follows. Firstly, the feature factors for the time-series data set are selected. Considering the structural features of LSTM and TCN, the input feature factors are numbered 1 to 24 in Table 2, which are forecast factors, time lagged variables and the longitude, latitude and altitude of the meteorological stations. Secondly, a time sliding window is applied to all time series (all features) at each station, and a number of batch samples with consistent length of time series will be generated by the time sliding window segmentation technique. Then, the batch samples generated at each station are connected to form the final data set. Among them, the setting of the time sliding window for the segmentation of time-series data is shown in Figure 2. For the feature data set (X in Figure 5), the length of the sliding window is set to 7 days, and after the sliding, the shape of the feature data samples in the training set is obtained as (1712 × 2238, 7, 24), where 1712 is the number of sequences generated at a single station after the time sliding window, and 2238 is the total number of stations. For the target data set (Y in Figure 5), the time-series length is set to 1, the day-by-day sliding is applied through the sliding window, and the shape of the target data sample is (1712 × 2238, 1, 1). So far, the feature data set and the target data set have achieved one-to-one correspondence on each batch. After completing the data set construction, the data set can be directly input into the LSTM and TCN.

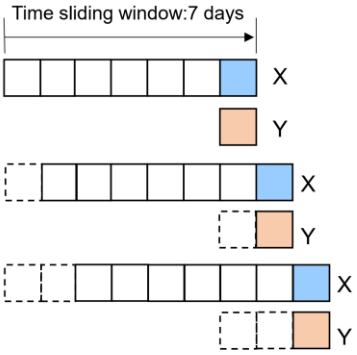

Figure 5. Schematic design of time sliding window in the construction of time-series data set.

3.2. Hyperparameter Settings and Training

For all neural network models (the ED-FCNN, LSTM, TCN and FLT), the common hyperparameters are as follows. The optimizer algorithm in the neural network is the adaptive moment estimation (Adam), and the batch size is set to 64. The number of neurons in the input layer and output layer is determined by the number of features in the input and output neural networks, respectively. The number of neurons in output layer of all model in this study is 1, and the activation function is the Linear function (linear units).

Specifically, the number of neurons in the hidden layer of the neural network in this study is determined by grid search. The number of neurons in the hidden layer is generally set to multiples of 16 and 16, such as 16, 32, 64, etc. Hence, the number of neurons in the hidden layer of the above neural networks is determined by means of grid search. The values of 16, 32, 64, 128 and 256 are set for the grid search, and a cross-validation method adaptive to time series is applied for the grid search, which can avoid the crossover of temporal features in the data, aiming to make the model selection steadier and more robust. Finally, the hyperparameters with the smallest average RMSE of all validation sets are selected. The results of hyperparameters selection of each neural network are as follows: The number of hidden layers is 1, the number of neurons in the hidden layer is 64 and 128 both in ED-FCNN and LSTM, respectively. The network structure of TCN is different from ED-FCNN and LSTM. TCN migrates Convolutional Neural Networks (CNN) to time series. Due to its specific network structure, the hyperparameters of TCN are different from those of ED-FCNN and LSTM. The main hyperparameters of TCN include convolution kernel size, expansion coefficient and convolution kernel number. Therefore, the hyperparameter setting of TCN in this study is: the size of one-dimensional convolution kernels is set as 3, and the dilations are set as 1, 2, 4, 8, 16 and 32, successively, and the number of convolution kernels of TCN is finally determined to be 128.

When training the neural network, the loss function is set to mean absolute error (MAE). Early stopping is selected as the regularization method to prevent the model from over-fitting. The number of model epochs is set to 50 times. If the validation error in training increases instead of decreasing for three consecutive times, the training is stopped and the model with the best performance is stored in the validation set.

3.3. Decaying Average Model

As a contrast, we compare performance of the proposed method with that of the decaying averaging (DA) method, which is an adaptive error correction method of KF type [37–39]. The principle of the DA method for error correction is as follows.

$$B(t) = (1 - w) \times B(t - 1) + w \times (F - a) \tag{1}$$

where $B(t)$ is the lagged average error of temperature forecasts for each leading time t at the station, and $B(t{-}1)$ is the lagged average error of the day before. F, a and w are the station forecast, observation and weight coefficient, respectively.

By referring to a previous study [40], the training period is set to 35 days in this study. In the training set, the method of day-by-day rolling update will be adopted at each station, and the weight coefficient corresponding to the minimum average RMSE at all stations is finally selected as the best weight of the corresponding station through comparisons among multiple groups of sensitivity tests. Here, the weight coefficient is set in the range of 0.001–1, with the step size being 0.001.

For a given station and leading time t, the steps of temperature forecast correction are as follows. Firstly, the cold start is carried out when t = 1, that is, $B(t - 1)$ = 0. Secondly, the optimal weight coefficient of W in the training period is calculated. Thirdly, the average lagged error $B(t)$ is calculated according to Equation (1). The fourth step is to repeat the third step. After the iterative accumulation during the training period, the obtained error tends to be stable and can represent the situation of systematic error to a certain extent. Finally, the corrected forecast value is obtained by subtracting the latest $B(t)$ from the current forecast.

3.4. Verification Method

The forecast performance for daily maximum temperature of each model is evaluated by using the RMSE, mean absolute error (MAE), the accuracy of temperature forecast (ATF).

$$\text{RMSE} = \sqrt{\frac{\sum_1^n (f(x_n) - y_n)^2}{n}} \tag{2}$$

$$\text{MAE} = \frac{1}{N}\sum_1^n |f(x_n) - y_n| \tag{3}$$

where f is the deep-learning-based regression function, x_n is the input, $f(x_n)$ is the model forecast, y_n is the observation, and n is the total number of forecast samples.

ATF is the percentage of the samples (Nr) whose absolute difference between the forecast value and observed value is no greater than m °C in the total samples (Nf) [19], where m is the temperature threshold. ATF is calculated as follows.

$$\text{ATF} = \frac{N_r}{N_f} \times 100\% \tag{4}$$

Following previous studies [19,41], m is set to 2 °C in this study, and the higher the ATF value is, the more accurate temperature forecast will be.

4. Results

4.1. General Evaluation

For each neural network model, DA and the ECMWF-IFS model, the validation metrics averaged over all stations during the entire validation period in 2020 are summarized in Table 3. Overall, for the test period in 2020, the neural network models (ED-FCNN, LSTM, TCN and FLT) and the DA reduce the average RMSE (MAE) by 43.27% to 52.36% (42.00% to 51.50%) and increase the average ATF by 34.12% to 43.12% compared with the ECMWF-IFS model output. Among them, only neural network models decreased the average RMSE (MAE) by 48.00% to 52.36% (46.50% to 51.50%) and improved the average ATFs by 38.89%

to 43.12%. The neural network models also reduce the average RMSEs (MAEs) by 8.33% to 16.03% (7.76% to 16.38%) and increase the average ATFs by 3.56% to 6.71% compared with the DA model. The FLT model integrated by the stacking ensemble technique achieves the best results, with the RMSE reduced by 52.36% compared with the ECMWF-IFS model and by 16.03% compared with the DA model, respectively. For the neural network model, the models of LSTM and TCN adaptive to the time-series data set perform better than the ED-FCNN model, but the ensemble learning model of FLT is better than the above three models. Compared with the ED-FCNN, LSTM and TCN models, the RMSEs of the FLT model are decreased by 8.39%, 1.50% and 2.96%, the MAEs are decreased by 9.35%, 3.00% and 3.00%, and ATFs are increased by 3.04%, 0.67% and 0.83%, respectively, indicating that the FLT model can combine the advantage of a single model to effectively reduce forecast error.

Table 3. Mean verification metrics of each neural network model for daily maximum temperature on the test set.

Metrics	ECMWF-IFS	ED-FCNN	LSTM	TCN	FLT	DA
RMSE (°C)	2.75	1.43	1.33	1.35	1.31	1.56
MAE (°C)	2.00	1.07	1.00	1.00	0.97	1.16
ATF (%)	62.04	86.17	88.2	88.06	88.79	83.21

In addition, the overall performance of all neural network models (the ED-FCNN, LSTM, TCN and FLT) on extreme events are also compared with the ECMWF-IFS model output (Table 4). Here, the extreme events are defined as those with the daily maximum temperature on the second day observed at the station being above the 90th percentile or below the 10th percentile during the study period (test set). The performance improvement of all neural network models compared with the ECMWF-IFS model on extreme events is similar to the overall evaluation. For samples of extreme events at the station with the daily maximum temperature above (below) the 90th (10th) percentile, compared with the ECMWF-IFS models, the RMSEs are improved by 60.63% to 66.67% (35.63% to 44.06%), the MAEs are decreased by 60.08% to 67.08% (33.15% to 41.57%) and the ATFs are increased by 81.96% to 89.92% (19.40% to 25.12%), respectively. The TCN (FLT) model has the largest improvement, followed by the FLT (TCN) model, for the extreme events above (below) the 90th (10th) percentile.

Table 4. Mean verification metrics for extreme events related to maximum temperature on the test set for each neural network model.

Metrics	ECMWF-IFS	ED-FCNN	LSTM	TCN	FLT	DA
RMSE (°C)	3.15 (2.61)	1.19 (1.68)	1.24 (1.49)	1.05 (1.47)	1.08 (1.46)	1.27 (1.77)
MAE (°C)	2.43 (1.78)	0.93 (1.19)	0.97 (1.07)	0.80 (1.06)	0.83 (1.04)	0.97 (1.27)
ATF (%)	49.50 (69.28)	91.31 (82.72)	90.07 (85.77)	94.01 (86.43)	93.39 (86.68)	89.29 (80.25)

The improvement of each neural network model is better than that of the DA model. For the samples of extreme events higher (lower) than the 90th (10th) percentile, the neural network models reduced the RMSEs by 2.36% to 17.32% (5.08% to 17.51%), the MAEs are decreased by 0% to 17.53% (6.30% to 18.11%) and ATFs are increased by 0.87% to 5.29% (3.08% to 8.01%) compared with the DA model. The improvement for samples lower than the 10th percentile is not as good as that of the samples above the 90th percentile in each neural network model.

4.2. Evaluation of the Spatio-Temporal Continuous Forecast Skill

Figure 6 shows the spatial distributions of RMSEs over the Chinese mainland for the daily maximum air temperature forecasts of different neural work models, the ECMWF-IFS and DA model. Great improvements are achieved by all neural network models and the DA compared with the ECMWF-IFS model. More than 94.33% stations of each neural network model have RMSE less than 2.0 °C, while the ECMWF-IFS has only 53.40% stations with RMSE less than 2.0 °C. The most obvious correction effect is mainly distributed in the Qinghai-Tibet Plateau, and the RMSE of ECMWF-IFS in this region is decreased from more than 3.0 °C to 2.0 °C.

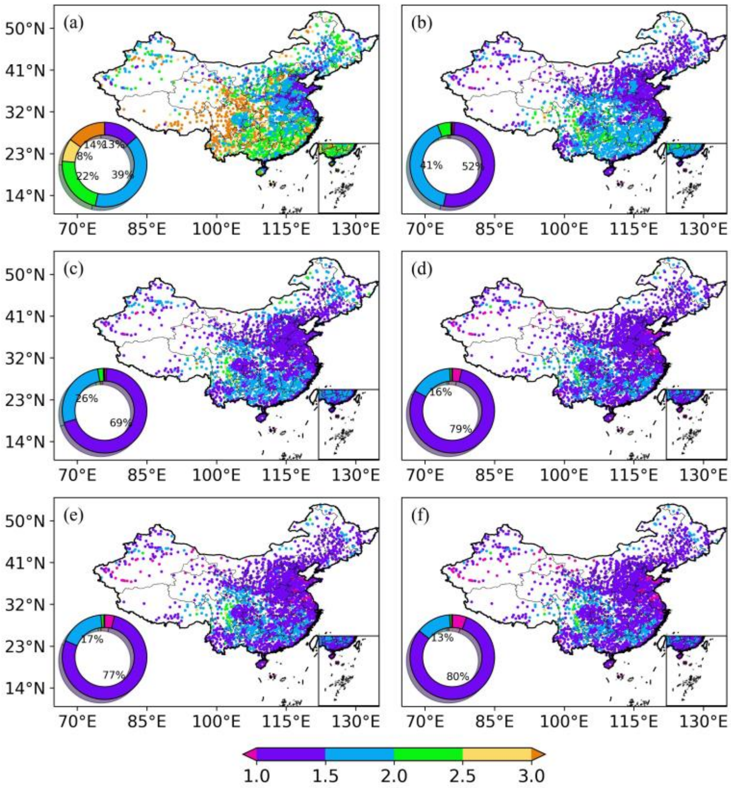

Figure 6. Spatial distributions of average RMSE of daily maximum air temperature forecasts at all stations using different neural work models and the ECMWF-IFS model (unit: °C; the ring chart in the lower left corner represents the proportion of the number of stations (less than 6% are not shown) with different RMSE interval values to the total number of stations). (**a**) ECMWF-IFS, (**b**) DA, (**c**) ED-FCNN, (**d**) LSTM, (**e**) TCN and (**f**) FLT.

Compared with the ECMWF-IFS and DA, the neural network models of ED-FCNN, LSTM, TCN and FLT, increase the proportion of stations with the RMSE between 1.0 °C and 1.5 °C from 52.37% of DA model and 13.67% of ECMWF-IFS to 69.53%, 79.00%, 77.35% and 79.94% respectively. Especially in the area south of the middle and lower reaches of the Yangtze River (See Figure 1 for the location), the RMSEs of most stations of the DA model are between 1.5 °C and 2.0 °C, while the RMSEs of all neural network models in this area are between 1.0 °C and 1.5 °C.

The spatial distributions of RMSEs of neural network models reveal that the proportions of stations with RMSE values less than 1.0 °C are 0.80%, 3.66%, 3.75% and 5.50% in ED-FCNN, LSTM, TCN and FLT, respectively. Thus, the FLT model shows the best performance in terms of the overall RMSE at stations. Moreover, compared with other neural network models, the RMSE in the Guangxi Province and Hunan Province (See Figure 1 for the location) is obviously reduced by the FLT model. This implies that the ensemble neural network integrated by the stacking of ED-FCNN, LSTM and TCN improves the forecast error based on the models with good performance in the Level-0 layer.

Figure 7 shows the daily RMSE of each model averaged at all stations in the test set. Among all the models, RMSE of the ECMWF-IFS model is much larger than those of other neural network models or the DA model, and fluctuates more greatly over time compared with other models, while RMSEs of the DA and other neural network models are relatively more stable over the year. The RMSE of each neural network models and DA in summer is lower than that in winter. The DA performs better in reducing the RMSE than the ECMWF-IFS, but each neural network model is even better. The average RMSE of four neural network models is 1.34 °C, and is 12.95% lower than that of DA model. The FLT model exhibits the lowest RMSE among the three single machine-learning models. Specifically, the FLT model maintains the lowest RMSE on a daily basis throughout the test period. That is, the FLT model yields the best performance compared with other models. FLT reduces the RMSE by 16.23% and 52.74% compared with the DA and ECMWF-IFS for the day-by-day average, respectively.

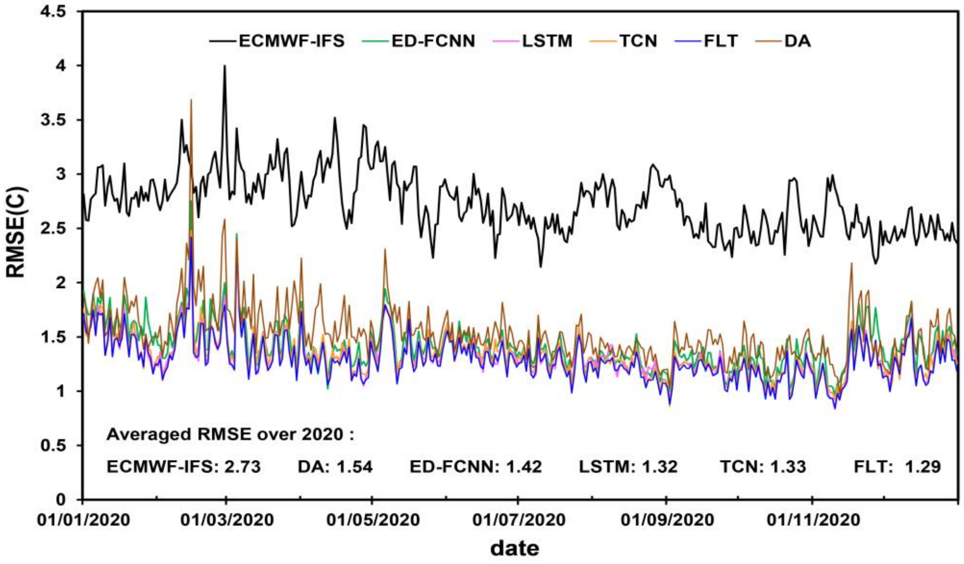

Figure 7. The average RMSE at all stations over the Chinese mainland on each calendar day during the validation period using five post-processing models and the ECMWF-IFS model (the average RMSE of each model is also shown in the top; unit: °C).

4.3. Accuracy of Maximum Air Temperature Forecast

The bivariate distributions of the forecast accuracy of daily maximum temperature at all stations by the ECMWF-IFS model and neural network models are given in Figure 8. The point falling on the upper left side of the diagonal line in the figure means that the forecast accuracy of daily maximum temperature at that station is greater than the forecast accuracy by the ECMWF-IFS model. As is shown in Figure 8, except for the ED-FCNN and DA models that have 19 and 47 cases of sample points located at the bottom right of the diagonal line, the rest of the models have less than 4 cases of sample points located at the bottom right of the diagonal line. That is, the majority of sample points fall at the upper left of the diagonal line for both neural network models and the DA model, indicating that the ATFs at almost all station have been improved compared to that by the ECMWF-IFS.

From the histogram above each main figure, the distribution of ATF in the ECMWF-IFS mode is a significantly negative skewness with the left tail being very long. The values of ATF are mainly concentrated between 75% and 85%. There are many sample points with an ATF value of less than 60%, and a considerable number of sample points with an ATF value of 0%. To the sample with the ATF between 0 and 60% in the ECMWF-IFS model have disappeared, and the value of ATF increases to more than 60% and even up to 90% for the vast majority of samples after the correction of each neural network model. The ATF values of the DA model are in the range of 60% to 95%, compared with the DA model, while the ATF values of each neural network model are in the range of 70% to 95%. It is significantly higher and more concentrated in the range of 85% to 95%. The LSTM model has the ATF distribution very close to the FLT model, while FLT performs the best among all the models.

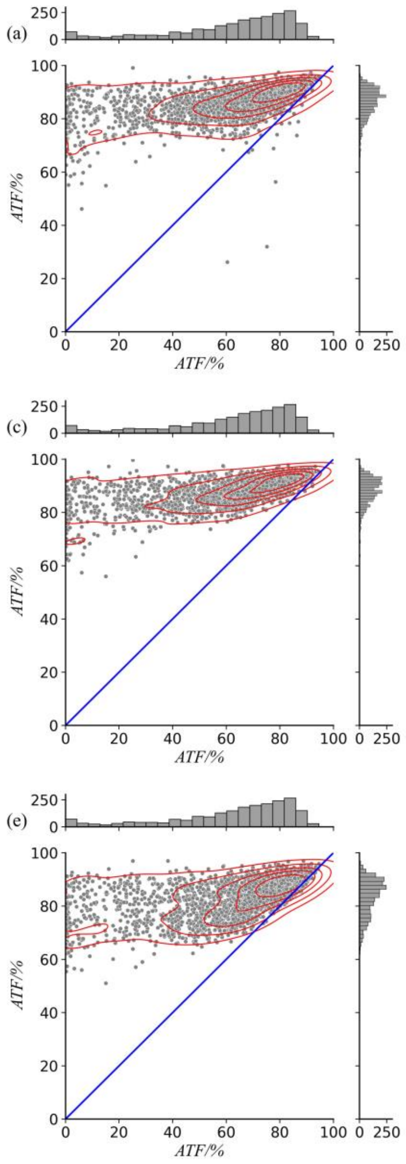

Figure 8. Bivariate of temperature forecast accuracy for (**a**) ED-FCNN, (**b**) LSTM, (**c**) TCN, (**d**) FLT and (**e**) DA model relative to the ECMWF-IFS model (horizontal coordinates are the ATF of the daily maximum temperature by the ECMWF-IFS model on the test set, vertical coordinates represent that of neural network models and DA model, red solid lines represent the contour of temperature forecast accuracy, blue solid lines are diagonal lines, the histogram at the top of the main figure is the distribution of temperature forecast accuracy of the ECMWF-IFS model, and the histogram at the right of the main figure is the frequency of temperature forecast accuracy of each neural network model).

5. Conclusions and Discussion

In this study, an ensemble deep neural network model named FLT is developed to forecast the daily maximum temperature over the Chinese mainland by using a stacking generalization approach, and its performance is further compared with three individual neural network models (i.e., ED-FCNN, LSTM and TCN) and a decaying averaging model (i.e., the DA method). To train this model, we use the daily maximum temperature data from 2238 national surface meteorological stations and forecast products from the ECMWF-IFS model in the past 4 years to validate the training effects of the FLT, the 3 individual neural network models and the decaying averaging model. The main findings are summarized as follows.

Each neural network model can effectively improve the forecast bias of the ECMWF-IFS model. Specifically, the RMSE (MAE) decrease is between 48.00% and 52.36% (46.50% and 51.50%) and the ATF increase is between 38.89% and 43.12% compared with the ECMWF-IFS model. In addition, the models also exhibit good performance for extreme events. For the extreme temperature events above the 90th percentile (or below the 10th percentile), the RMSE decrease ranges from 60.63% to 66.67% (from 35.63% to 44.06%) and the ATF increase ranges from 81.96% to 89.92% (from 19.40% to 25.12%) compared with the ECMWF-IFS model. Among them, the FLT model yields the best performance compared with other models.

The neural network models can improve the forecast accuracy at almost all stations. All neural models can significantly improve the prediction ability of samples from the ECMWF-IFS model, whose ATFs are less than 60%. The ATFs of each neural network model are significantly concentrated in a relatively high level (85% to 95%) compared with the DA model. In addition, all models can significantly improve the forecast performance of the ECMWF-IFS model in day-by-day, and the forecast performance is better in summer than in winter. This is probably because the cold air activity is stronger and more frequent in winter than in summer, and the forecast ability of neural network models is weaker when the temperature changes abruptly. Compared with the commonly used DA model (1.54), FLT (1.29) improves the day-by-day averaged RMSE by 16.23%.

Different neural network architectures have different adaptability to similar modeling tasks. In this study, the learning effect and forecast performance of the LSTM and the TCN are better than the ED-FCNN. Although two features of month and season that characterize the temporal information have been added to the data samples for the ED-FCNN, the LSTM and the TCN with the input of time-series data set are able to mine the time dependencies in the time-series data, indicating that the construction of the time-series data set is crucial. The ensemble learning model constructed by applying stacked generalization based on different types of neural networks can effectively reduce the forecast bias in the Level-0 model and enhance the model's generalization ability.

However, some issues in this study still remain to be explored in depth.

For example, the bilinear interpolation method was used to interpolate the forecast data to the stations during the data preprocessing, which inevitably brought some errors. Additionally, the physical quantities input into the model were not in the form of a grid, which might destroy the relatively complete structure of the surface pressure field and the structure of the upper-air synoptic situation field. In future studies, it is suggested that the grid structure of the forecast products can be retained to construct new models based on the new forms of input and output data (grid as input and station as output in this study). In addition, the extraction of spatio-temporal features of meteorological data can also be achieved by fusing the two-dimensional convolutional networks with the RNNs. As to feature selection, this study adopts the combination of variance filtering, correlation analysis and mutual information value, and the effect of other feature selection methods on model effectiveness can be explored in future research.

In terms of model construction, this study selects the size of 7 days for the time window when constructing the time-series data set, while there were no more sensitivity tests on the size of time window. Because the LSTM unit is relatively sensitive to the time-series length, which usually cannot handle the long-range dependence, the time-series length may have some influence on the modeling effect. In addition, this study only compares the effects of different models on modeling effects and explores the ability of different types of neural network frameworks in dealing with the problem of error correction for a model's temperature forecast. The reasons for how features affect neural networks to make them work have yet to be explored, which is also limited by the uninterpretability of neural networks themselves. This may be the focus of future research work.

In general, the prediction performance of the decreasing average method is between ECMWF-IFS and the neural network, and the improvement of ECMWF-IFS is significant, but it is far less than that of the single neural network, let alone the ensemble neural network model. The deep learning, including mentioned in this paper, belongs to the MOS method. Compared with the MOS method, the deep learning method can extract more complex nonlinear relationships between features and targets. The results of this study show that the ensemble learning methods based on stacked generalization can obtain better prediction results than traditional statistical post-processing methods and single deep learning methods. However, ensemble learning methods are also diverse, and the effects of ensemble methods such as Snapshot Ensembling and Boosting need to be further tested. In addition, this study only verifies the effectiveness of each model to correct the forecast error of daily maximum temperature of the ECMWF-IFS model at the national meteorological stations, and the applicability to other forecast leading time and other meteorological elements (e.g., precipitation) needs to be further verified. Moreover, the effectiveness of their performance on gridded data can also be measured.

Author Contributions: Conceptualization, L.Z. and S.L.; data curation, S.L. and L.Z.; formal analysis, S.L. and L.Z.; visualization, S.L.; funding acquisition, L.Z.; investigation, S.L., D.Q. and L.Z.; methodology, L.Z., S.L. and D.Q.; writing—original draft, S.L. and L.Z.; writing—review and editing, L.Z., S.L. and D.Q. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Science and Technology Development Fund of Chinese Academy of Meteorological Sciences (2023KJ022) and Major Science and Technology Program for Water Pollution Control and Treatment (2013ZX07304-001-1).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We would like to acknowledge the reviewers and editor for their helpful comments and suggestions.

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

References

1. Blunden, J.; Arndt, D.S. State of the Climate in 2019. *Bull. Am. Meteorol. Soc.* **2020**, *101*, Si-S429. [CrossRef]
2. Verkade, J.S.; Brown, J.D.; Reggiani, P.; Weerts, A.H. Post-processing ECMWF precipitation and temperature ensemble reforecasts for operational hydrologic forecasting at various spatial scales. *J. Hydrol.* **2013**, *501*, 73–91. [CrossRef]
3. Anderson, B.G.; Bell, M.L. Weather-related mortality: How heat, cold, and heat waves affect mortality in the United States. *Epidemiology* **2009**, *20*, 205–213. [CrossRef] [PubMed]
4. Tran, T.T.K.; Lee, T.; Shin, J.Y.; Kim, J.S.; Kamruzzaman, M. Deep learning-based maximum temperature forecasting assisted with meta-learning for hyperparameter optimization. *Atmosphere* **2020**, *11*, 487. [CrossRef]
5. Shen, X.; Wang, J.; Li, Z.; Chen, D.; Gong, J. Research and operational development of numerical weather prediction in China. *J. Meteorol. Res.* **2020**, *34*, 675–698. [CrossRef]
6. Kwon, I.; English, S.; Bell, W.; Potthast, R.; Collard, A.; Ruston, B. Assessment of progress and status of data assimilation in numerical weather prediction. *Bull. Am. Meteorol. Soc.* **2018**, *99*, ES75–ES79. Available online: https://journals.ametsoc.org/view/journals/bams/99/5/bams-d-17-0266.1.xml (accessed on 21 December 2022). [CrossRef]
7. Hamill, T.M.; Snyder, C.; Morss, R.E. A comparison of probabilistic forecasts from bred, singular-vector, and perturbed observation ensembles. *Mon. Weather Rev.* **2000**, *128*, 1835–1851. [CrossRef]
8. Hamill, T.M.; Snyder, C.; Whitaker, J.S. Ensemble forecasts and the properties of flow-dependent analysis-error covariance singular vectors. *Mon. Weather Rev.* **2003**, *131*, 1741–1758. [CrossRef]
9. Vislocky, R.L.; Fritsch, J.M. Performance of an advanced MOS system in the 1996-97 National Collegiate Weather Forecasting Contest. *Bull. Am. Meteorol. Soc.* **1997**, *78*, 2851–2858. [CrossRef]
10. Wilks, D.S.; Hamill, T.M. Comparison of ensemble-MOS methods using GFS reforecasts. *Mon. Weather Rev.* **2007**, *135*, 2379–2390. [CrossRef]
11. Klein, W.H.; Lewis, F. Computer forecasts of maximum and minimum temperatures. *J. Appl. Meteorol.* **1970**, *9*, 350–359. [CrossRef]
12. Glahn, H.R.; Lowry, D.A. The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteorol. Climatol.* **1972**, *11*, 1203–1211. [CrossRef]

13. Kalman, R.E. A new approach to linear filtering and prediction problems. *J. Basic Eng.* **1960**, *82*, 35–45. [CrossRef]
14. Wilson, L.J.; Beauregard, S.; Raftery, A.E.; Verret, R. Calibrated surface temperature forecasts from the Canadian ensemble prediction system using bayesian model averaging. *Mon. Weather Rev.* **2007**, *135*, 1364–1385. [CrossRef]
15. Hart, K.A.; Steenburgh, W.J.; Onton, D.J.; Siffert, A.J. An evaluation of mesoscale-model-based model output statistics (MOS) during the 2002 Olympic and Paralympic winter games. *Weather Forecast.* **2003**, *19*, 200–218. [CrossRef]
16. Stensrud, D.J.; Yussouf, N. Short-range ensemble predictions of 2-m temperature and dewpoint temperature over New England. *Mon. Weather Rev.* **2003**, *131*, 2510–2524. [CrossRef]
17. Rasp, S.; Lerch, S. Neural networks for postprocessing ensemble weather forecasts. *Mon. Weather Rev.* **2018**, *146*, 3885–3900. [CrossRef]
18. Men, X.; Jiao, R.; Wang, D.; Zhao, C.; Liu, Y.; Xia, J.; Li, H.; Yan, Z.; Sun, J.; Wang, L. A temperature correction method for multi-model ensemble forecast in north China based on machine learning. *Clim. Environ. Res.* **2019**, *24*, 116–124. [CrossRef]
19. Li, H.; Chen, Y.; Xia, J.; Wang, Y.; Zhu, J.; Zhang, P. A model output machine learning method for grid temperature forecast in the Beijing area. *Adv. Atmos. Sci.* **2019**, *36*, 1156–1170. [CrossRef]
20. Xia, J.; Li, H.; Kang, Y.; Yu, C.; Ji, L.; Wu, L.; Lou, X.; Zhu, G.; Wang, Z.; Yan, Z. Machine learning-based weather support for 2022 Winter Olympics. *Adv. Atmos. Sci.* **2020**, *37*, 927–932. [CrossRef]
21. Cho, D.; Yoo, C.; Im, J.; Cha, D.-H. 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]
22. Bai, S.; Kolter, J.Z.; Koltun, V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv* **2018**, arXiv:1803.01271. [CrossRef]
23. Hewage, P.; Trovati, M.; Pereira, E.; Behera, A. Deep learning-based effective fine-grained weather forecasting model. *Pattern Anal. Appl.* **2021**, *24*, 343–366. [CrossRef]
24. 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]
25. Chen, K.; Wang, P.; Yang, X.; Zhang, N.; Wang, D. A model output deep learning method for grid temperature forecasts in Tianjin area. *Appl. Sci.* **2020**, *10*, 5808. [CrossRef]
26. Zhao, L.; Lu, S.; Qi, D.; Xu, D.; Ying, S. Daily maximum air temperature forecastbased on fully connected neural network. *J. Appl. Meteorol. Sci.* **2022**, *33*, 257–269. (In Chinese) [CrossRef]
27. Lee, J.; Im, J.; Kim, K.; Quackenbush, L.J. Machine learning approaches for estimating forest stand height using plot-based observations and airborne LiDAR data. *Forests* **2018**, *9*, 268. [CrossRef]
28. Liu, T.; Abd-Elrahman, A.; Morton, J.; Wilhelm, V.L. Comparing fully convolutional networks, random forest, support vector machine, and patch-based deep convolutional neural networks for object-based wetland mapping using images from small unmanned aircraft system. *GISci. Remote Sens.* **2018**, *55*, 243–264. [CrossRef]
29. Park, S.; Im, J.; Park, S.; Yoo, C.; Han, H.; Rhee, J. Classification and mapping of paddy rice by combining Landsat and SAR time series data. *Remote Sens.* **2018**, *10*, 447. [CrossRef]
30. Wylie, B.K.; Pastick, N.J.; Picotte, J.J.; Deering, C.A. Geospatial data mining for digital raster mapping. *GISci. Remote Sens.* **2019**, *56*, 406–429. [CrossRef]
31. Belkin, M.; Hsu, D.; Ma, S.; Mandal, S. Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 15849–15854. [CrossRef] [PubMed]
32. Chou, J.-S.; Pham, A.-D. Enhanced artificial intelligence for ensemble approach to predicting high performance concrete compressive strength. *Constr. Build. Mater.* **2013**, *49*, 554–563. [CrossRef]
33. Healey, S.P.; Cohen, W.B.; Yang, Z.; Brewer, C.K.; Brooks, E.B.; Gorelick, N.; Hernandez, A.J.; Huang, C.; Joseph Hughes, M.; Kennedy, R.E. Mapping forest change using stacked generalization: An ensemble approach. *Remote Sens. Environ.* **2018**, *204*, 717–728. [CrossRef]
34. Ren, Y.; Zhang, L.; Suganthan, P.N. Ensemble classification and regression-recent developments, applications and future directions. *IEEE Comput. Intell. Mag.* **2016**, *11*, 41–53. [CrossRef]
35. Chen, Y.; Huang, X.; Li, Y.; Chen, R.; Xu, Z.; Huang, X. Ensemble learning for bias correction of station temperature forecast based on ECMWF products. *J. Appl. Meteorol. Sci.* **2020**, *31*, 494–503. [CrossRef]
36. Wolpert, D.H. Stacked generalization. *Neural Netw.* **1992**, *5*, 241–259. [CrossRef]
37. Cui, B.; Toth, Z.; Zhu, Y.J.; Hou, D.H. Bias correction for global ensemble forecast. *Weather Forecast.* **2012**, *27*, 396–410. [CrossRef]
38. Glahn, B.; National Weather Service Office of Science and Technology; Meteorological Development Laboratory. *Bias Correction of MOS Temperature and Dewpoint Forecasts (MDL Office Note. 12-1)*; U.S. Department of Commerce National Oceanic and Atmospheric Administration: Washington, DC, USA, 2012. Available online: https://repository.library.noaa.gov/view/noaa/6913 (accessed on 21 December 2022).
39. Dube, A.; Singh, H.; Ashrit, R. Heat waves in India during MAM 2019: Verification of ensemble based probabilistic forecasts and impact of bias correction. *Atmos. Res.* **2020**, *251*, 105421. [CrossRef]

40. Cui, B.; Toth, Z.; Zhu, Y.J.; Hou, D.; Unger, D.; Beauregard, S. The trade-off in bias correction between using the latest analysis/modeling system with a short, versus an older system with a long archive. In Proceedings of the First THORPEX International Science Symposium, World Meteorological Organization, Montréal, QC, Canada, 6–10 December 2004; pp. 281–284.
41. Xiong, M. Calibrating daily 2 m maximum and minimum air temperature forecasts in the ensemble prediction system. *J. Meteorol. Res.* **2017**, *75*, 211–222. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.