*atmosphere*

MDPI

*Article*

# Prediction Skill of Extended Range 2-m Maximum Air Temperature Probabilistic Forecasts Using Machine Learning Post-Processing Methods

**Ting Peng [1,2], Xiefei Zhi [1,2,*], Yan Ji [1,2], Luying Ji [1,2] and Ye Tian [1]**

[1] Collaborative Innovation Center on Forecast and Evaluation of Meteorological Disasters (CIC-FEMD)/Key Laboratory of Meteorological Disasters, Ministry of Education (KLME), Nanjing University of Information Science & Technology, Nanjing 210044, China; pengting921@163.com (T.P.); 20171201013@nuist.edu.cn (Y.J.); luyingji@nuist.edu.cn (L.J.); tianye@nuist.edu.cn (Y.T.)

[2] WeatherOnline Institute of Meteorological Applications, Wuxi 214000, China

* Correspondence: zhi@nuist.edu.cn; Tel.: +86-186-5181-4986

Received: 9 June 2020; Accepted: 28 July 2020; Published: 4 August 2020


**Abstract:** The extended range temperature prediction is of great importance for public health, energy and agriculture. The two machine learning methods, namely, the neural networks and natural gradient boosting (NGBoost), are applied to improve the prediction skills of the 2-m maximum air temperature with lead times of 1–35 days over East Asia based on the Environmental Modeling Center, Global Ensemble Forecast System (EMC-GEFS), under the Subseasonal Experiment (SubX) of the National Centers for Environmental Prediction (NCEP). The ensemble model output statistics (EMOS) method is conducted as the benchmark for comparison. The results show that all the post-processing methods can efficiently reduce the prediction biases and uncertainties, especially in the lead week 1–2. The two machine learning methods outperform EMOS by approximately 0.2 in terms of the continuous ranked probability score (CRPS) overall. The neural networks and NGBoost behave as the best models in more than 90% of the study area over the validation period. In our study, CRPS, which is not a common loss function in machine learning, is introduced to make probabilistic forecasting possible for traditional neural networks. Moreover, we extend the NGBoost model to atmospheric sciences of probabilistic temperature forecasting which obtains satisfying performances.

**Keywords:** machine learning; probabilistic temperature forecast; extended range; neural networks; natural gradient boosting

## 1. Introduction

Subseasonal or extended range weather forecasts are used to predict heat waves, extreme cold events, thunderstorms, droughts and floods as far as four weeks ahead. Subseasonal forecasts can deliver relevant weather information, such as the timing of the onset of a rainy season and the risk of extreme rainfall events or heat waves. However, there is a well-known gap in current numerical prediction systems for the subseasonal timescale of 10 days to one month. This gap falls between medium-range weather forecasts (up to 10 days) and seasonal climate predictions (longer than one month). Medium-range weather forecasts are influenced by the initial conditions of the atmosphere, whereas predictions of the seasonal climate are more affected by slowly evolving surface boundary conditions, such as the sea surface temperature and soil moisture content [1–6]. Predictions on the subseasonal timescale have made progress in some regions and seasons [1,7,8], although the full potential of their predictability requires further exploration.

The Subseasonal Experiment (SubX) [9] is a multi-model ensemble experiment for subseasonal prediction. It is a research-to-operations project including global climate prediction models from both operation and research centers. It was developed to provide guidance for real-time subseasonal prediction and to improve forecast skills. SubX covers seven global models from US and Canadian modeling groups and has produced 17 years of historical retrospective (re)forecasts. It is an open research database (http://iridl.ldeo.columbia.edu/SOURCES/.Models/.SubX/) designed to operational standards [9]. The reforecasts from all seven global models are required to include, but are not limited to, the period from 1999 to 2015, with at least three ensemble members, a minimum of weekly initialization and forecasting at least 32 days in advance. In addition, some SubX models have also provided more than 18 months of real-time forecasts since 2016.

The systematic bias of the raw ensemble forecasts can be corrected using statistical post-processing methods. These methods reduce bias by learning a function that relates predictors to the variable of interest, which can be viewed as a supervised machine learning task. Bayesian model averaging [10] and ensemble model output statistics (EMOS) [11] are two state-of-the-art methods used in probabilistic forecasting. However, both of these methods specifically rely on parametric forecast distributions, which means that a predictive distribution has to be specified in advance and its parameters estimated. We have used the EMOS frame as a benchmark method because of its superior performance [11] and time saving capability. Two alternative approaches based on machine learning, neural networks and natural gradient boosting (NGBoost), which learn nonlinear mapping using abundant predictors in a data-driven way, are also considered here.

Compared to the traditional multi-model superensemble forecasts which may perform better than the best individual model forecasts and multi-model ensemble mean forecasts [12–16], a neural network is a flexible machine learning algorithm that can deal with complex problems using arbitrary nonlinear functions [17]. A neural network consists of interconnected nodes organized in layers and regulated by an activation function. Neural networks are used in a number of different fields, including computer vision and natural language processing [18], as well as in biology, physics and chemistry [19,20]. In the atmospheric sciences, neural networks have been applied to precipitation nowcasting [21,22] as well as short- and medium-range weather forecasts [23]. Rasp and Lerch [24] trained neural networks without a hidden layer and with a single hidden layer for post-processing ensemble temperature forecasts to establish nonlinear mapping between the outputs of numerical models and the corresponding observations to reduce the bias from raw ensemble prediction systems.

NGBoost is an algorithm that allows gradient boosting to make probabilistic forecasts in a generic way [25]. Gradient boosting is a supervised learning method that combines several weak learners to give an additive ensemble [26]. The gradient boosting method has been widely applied in prediction tasks, although it has rarely been applied to probabilistic forecasting. NGBoost combines the natural gradient with a multi-parameter boosting algorithm to estimate intuitively how parameters vary with the observed features. Experiments on several regression datasets have shown that NGBoost is more flexible, modular and faster than existing methods for probabilistic forecasting [25,27]. We used two post-processing methods (neural networks and NGBoost) in a machine learning framework and the EMOS benchmark post-processing method to extend the range of the 2-m maximum air temperature probabilistic forecast.

This paper is organized as follows. Section 2 describes the datasets used, followed by the introduction of three post-processing methods and scoring rules in Section 3. Section 4 presents our main results and a summary is provided in Section 5. The discussions about the two machine learning methods are given in Section 6.

## 2. Data

We obtain raw ensemble forecasts from the Environmental Modeling Center, Global Ensemble Forecast System (EMC-GEFS) model under the SubX project (http://iridl.ldeo.columbia.edu/SOURCES/.Models/.SubX/) [9]. The base model of the EMC-GEFS is a numerical weather prediction

atmosphere–land model forced with prescribed sea surface temperatures. This base model contributes 11 ensemble members to the SubX reforecasts. The group provides reforecasts for the 1999–2016 period with weekly initialization and a lead time of 1–35 days. The anomaly correlation coefficients of precipitation and the 2-m temperature for week three, the anomaly correlation coefficients of the Real-time Multivariate Madden–Julian Oscillation indices and the North Atlantic Oscillation index all prove the superior prediction skills of the EMC-GEFS model [9].

We focus on the 2-m maximum air temperature forecasts over East Asia (15–60° N, 70–140° E) with a resolution of 1° × 1° for a lead time of 1–35 days. To ensure that our verification procedure mimics operational conditions, we set aside the data for the year 2016 as a validation set. The observations used for verification are obtained from the National Oceanic and Atmospheric Administration Climate Prediction Center (ftp://ftp.cdc.noaa.gov/Datasets/cpc_global_temp/) [9]. For the 2-m maximum air temperature over the land, the Climate Prediction Center provides a maximum daily temperature ($T_{\max}$) dataset with a horizontal resolution of 0.5° × 0.5° [28]. The verification data are re-gridded to a coarser EMC-GEFS model resolution of 1° × 1°.

## 3. Post-Processing and Verification Methods

### *3.1. Ensemble Model Output Statistics*

EMOS is a variant of the model output statistics method and regression techniques designed for probabilistic forecasting [29,30]. In the simple frame of model output statistics, which only uses the ensemble member forecasts $\boldsymbol{x}$ or $x_1, \ldots, x_k$ as predictors and multiple linear regression as the transfer function, the predictand $y$ can be written as:

$$y = a + b_1 x_1 + \cdots + b_K x_K \tag{1}$$

where $a$ and $b_1, \ldots, b_k$ (or denoted by $\boldsymbol{b}$) are the regression coefficients; $K$ is the number of ensemble members.

There has been little research into the application of regression techniques to probabilistic forecasting [31]. Following Gneiting et al. [11], the conditional distribution of the predictand $y$ based on the ensemble member forecast $\boldsymbol{x}$can be modeled by a single parametric distribution $P_\theta$with parameters $\theta \in \mathbb{R}^d$:

$$y|\boldsymbol{x} \sim P_\theta(\boldsymbol{x}) \tag{2}$$

When the distribution of the weather variable of interest $y$ (e.g., temperature) is Gaussian, Equation (2) can be written as:

$$y|\boldsymbol{x} \sim \mathcal{N}\left(\mu, \sigma^2\right) \tag{3}$$

where $\mu$ is the mean and $\sigma$ is the standard deviation. By applying regression theory to probabilistic forecasting, the EMOS method attempts to reduce the bias between the predictive mean/variance and the regression estimate by using a bias-corrected weighted average. Hence we use a linear function with the ensemble mean and spread to fit the predictive mean and variance:

$$\begin{cases} \mu = a + b_1 x_1 + \cdots + b_K x_K \\ \sigma^2 = c + dS^2 \end{cases} \tag{4}$$

where $S^2$ is the ensemble spread and $c$ and $d$ are nonnegative coefficients. Combining Equations (3) and (4), the Gaussian predictive distribution can be written as:

$$y|\boldsymbol{x} \sim \mathcal{N}\left(a + b_1 x_1 + \cdots + b_K x_K, c + dS^2\right) \tag{5}$$

These EMOS coefficients $\theta$= ($a$, $\boldsymbol{b}$, $c$, $d$) are estimated by minimizing the correct scoring rule (e.g., the continuous ranked probability score (CRPS) or the maximum likelihood estimation

(MLE)) during the training period. Our EMOS experiments are implemented in R using the scoringRules package [32].

### 3.2. Neural Networks

As in the flowchart of neural networks shown in Figure 1, the neural networks consist of nodes organized in layers. The first (input) layer contains the input features, whereas the last (output) layer represents the output targets. The layers between the input and output layers are referred to as hidden layers. Apart from the input features, each node in the network can be computed as:

$$z_j^l = f\left(a_j^l\right) \tag{6}$$

where $f$ is an activation function, $z_j^l$ is the value of the $j$th node in the $l$th layer and $a_j^l$ is the weighted average of the outputs $z_i^{l-1}$ in the previous layer. Additionally, $a_j^l$ can be written as: $a_j^l = \sum_{i=1}^{m^l} w_{ji}^l z_i^{l-1} + b$; $w_{ji}^l$ is the weight between the $i$th node in the $(l-1)$th layer and the $j$th node in the $l$th layer; $b$ is a bias term and is usually constant. The activation function $f$ is usually a nonlinear function which allows the network to be more complex and robust and the rectified linear unit (ReLU) is used here apart from in the output layer. The weights and biases are optimized to reduce the loss function using the Adam optimization method [33]. The topological structure of the neural networks is of great importance for its performance in which the number of nodes in each hidden layer is usually optimized via cross-validation. In our study, we take the 11 raw ensemble forecasts as the inputs and the predictive parameter μ and σ as the targets of the neural networks. Then we use an analogous-linear search method to test the optimal configuration of the neural networks where we build several models which consist of different number of nodes (i.e., 8, 16, 32, . . . , 256, 512, 1024) in the hidden layers. After the assessments (not shown), we finally build the neural networks model consisting of two hidden layers which contains 64 and 256 nodes, respectively.

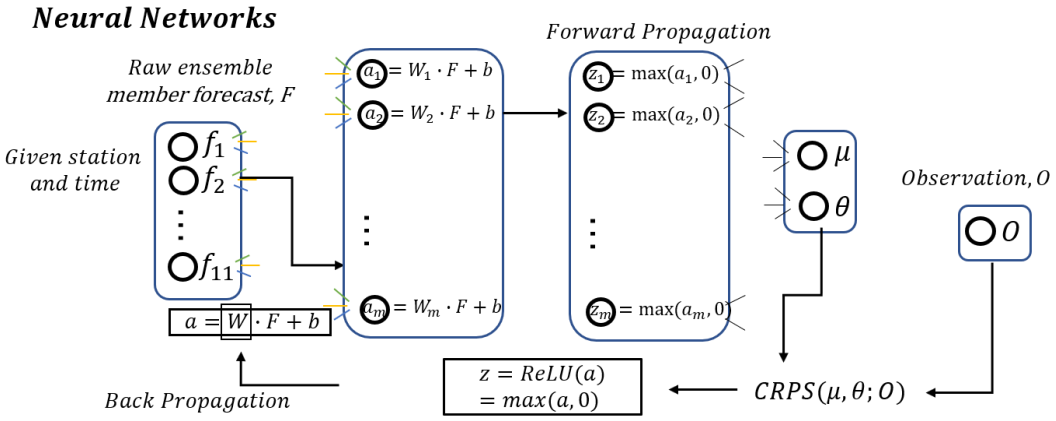


**Figure 1.** The flowchart of neural networks. $\boldsymbol{F}$ or $f_1, f_2, \ldots, f_{11}$, the raw ensemble forecasts and as the inputs; $\boldsymbol{W}$ or $W_1$, $W$, . . . , $W_m$, the weights of nodes; $\boldsymbol{a}$ or $a_1$, $a_2$, . . . ,$a_m$, the weighted average of the outputs in the previous layer; $\boldsymbol{z}$ or $z_1$, $z_2$, . . . , $z_m$, the outputs with the rectified linear function, ReLU, applied in $\boldsymbol{a}$; $\mu$ and $\sigma$, the two target parameters: mean and standard deviation; $0$, the verification or the observation; CRPS, the continuous ranked probability score. Here, $m$ is the number of the nodes.

Neural networks can be applied to a range of problems but have rarely been applied to probabilistic forecasting [24]. The difficulty lies in building the correct loss function for probabilistic forecasting. Here, we use a closed form expression of the CRPS (see Section 3.4) with a Gaussian distribution for temperature probabilistic forecasting. The experiment described the conditional distribution of the observation $y$ given the ensemble member forecast $\boldsymbol{x}$ as the input.

### 3.3. Natural Gradient Boosting

Figure 2 presents the flowchart of NGBoost. NGBoost is a supervised learning method for probabilistic forecasting. The approach uses the natural gradient to address the technical challenges that are difficult in generic probabilistic forecasts with existing gradient boosting methods. The origins of the natural gradient can be traced to the field of information geometry [34], where it was initially defined for the statistical manifold with the distance measure using Kullback–Leibler divergence [35]. The generalized natural gradient is the direction of steepest ascent in Riemannian space and is formed as:

$$\widetilde{\nabla} S(\theta, y) \propto \mathcal{I}_S(\theta)^{-1} \nabla S(\theta, y) \tag{7}$$

where $\mathcal{I}_S(\theta)$ is the Riemannian metric and $\nabla S(\theta, y)$ is the ordinary gradient of a scoring rule $S$. The natural gradient is invariant to parametrization, which distinguishes it from the ordinary gradient and helps to reflect how the space of distribution moves when updating. The algorithm has three main components: (1) the proper scoring rule (e.g., MLE or CRPS); (2) the parametric probability distribution (e.g., normal or Laplace); (3) the base learner (e.g., decision tree or linear). Different configurations are required for different kinds of problems.

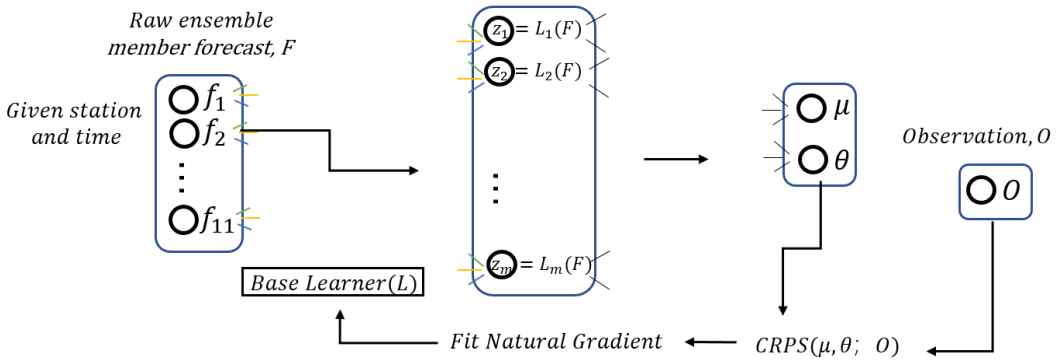


**Figure 2.** The flowchart of natural gradient boosting (NGBoost). $\boldsymbol{F}$ or $f_1, f_2, \ldots, f_{11}$, the raw ensemble forecasts and as the inputs; $\boldsymbol{L}$ or $L_1, L_2, \ldots, L_m$, the base learner; $\mu$ and $\sigma$, the two target parameters: mean and standard deviation; $0$, the verification or the observation; CRPS, the continuous ranked probability score. Here, $m$ is the number of base learners.

By assigning a score between a forecasted probability density $f$ at the verifying observation $y$, the scoring rule $S(f, y)$ represents the bias between the forecast and the true distribution [36]. Proper scoring rules prompt the model to obtain more calibrated probabilities during training as loss functions. The most commonly used proper scoring rules are MLE and CRPS, although CRPS is generally considered more robust than MLE [37]. Therefore, we chose CRPS as the target scoring rule in our experiments (see Section 3.4).

We focused on temperature probabilistic forecasting. The input $\boldsymbol{x}$ contains raw ensemble forecasts and the target $y$ is the corresponding observation. Assuming that the variable temperature follows a normal distribution, $\theta = (\mu, log\ \sigma)$ are appointed as the predicted parameters and linear learners are used here as the base learner $l$ to speed up the calculation. The total number $M$ of training linear learners is set up to 100.

To obtain the predicted parameters $\theta$for the input $\boldsymbol{x}$, each base learner $l^{(k)}$ $(k = 1, \ldots, m)$ first makes an independent prediction based on the same $\boldsymbol{x}$and then the results are integrated. Note that there are two base learners $l^{(k)} = \left(l_{\mu}^{(k)}, l_{log\ \sigma}^{(k)}\right)$ per stage for a normal distribution with parameters $\mu$ and $log\ \sigma$.

We use a stage-specific scaling factor $\rho^{(k)}$ and a common learning rate $\eta$ to scale the predicted outputs in $i$th iteration:

$$\theta^{(i)} = \theta^{(i-1)} - \eta \sum_{k=1}^{m} \rho^{(k)} \cdot l^{(k)}(\boldsymbol{x}) \tag{8}$$

The predicted parameters are updated to $\theta^{(i)}$ by adding to each $\theta^{(i-1)}$ the scaled projected natural gradient $\rho^{(k)} \cdot l^{(k)}(\boldsymbol{x})$ and scaled by a small learning rate η. The learning algorithm starts with a random common $\theta^{(0)}$ and minimizes the sum of the scoring rule $S$ by training all the training samples. More details are available in Duan et al. [25].

### *3.4. Verification Methods*

Probabilistic forecasting aims to maximize the sharpness of the predictive distributions through calibration [38]. For calibration, the forecast probability density functions (PDFs) and verifications are expected to be consistent with each other and, for sharpness, the prediction intervals are expected to be narrowed by post-processing methods to obtain a more concentrated and sharper forecast PDF.

The Talagrand diagrams, also known as the verification rank histogram [39–42] and the probability integral transform (PIT) histograms are often adopted to evaluate the calibration of the ensemble forecast. They are analogous, but the Talagrand diagrams tend to assess the spread, whereas the PIT histograms assess the PDF forecasts. An ideal ensemble forecast usually behaves as a uniform distribution. However, in most cases, it exhibits U-shaped PIT histograms due to the under-dispersive forecasts. Here, PIT histograms have been selected to discuss the calibration of the post-processing methods.

In addition to showing PIT histograms, we computed the coverage and average width of the 88.33% central prediction interval, which is chosen from the range of an 11-member ensemble. The coverage represents the accuracy while the average width is used to assess the sharpness.

To verify the deterministic forecasts, we computed the mean absolute error (MAE) and the root-mean-square error (RMSE). Denoting $\mu_{s,t}$ as a deterministic forecast and $y_{s,t}$ as the observation, the MAE is defined as:

$$\mathrm{MAE} = \frac{1}{N} \sum_{n=1}^{N} \left| y_n - \mu_n \right| \tag{9}$$

and the RMSE can be written as:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N} \sum_{n=1}^{N} (y_n - \mu_n)^2} \tag{10}$$

where $N$ is the number of cases in the training set.

We also computed the CRPS for the assessment of predictive PDFs to address the calibration and sharpness simultaneously. The CRPS is the integral of the Brier scores at all possible threshold values for a continuous predictand [43]. Denoting $F_\theta$ as the predictive cumulative distribution function (CDF) with parameters $\theta$ and $y$ as the observations, the CRPS is defined as:

$$\mathrm{crps}(F_\theta, y) = \int_{-\infty}^{y} F_\theta(z)^2 dz + \int_{y}^{\infty} (1 - F_\theta(z))^2 dz \tag{11}$$

when $F_\theta$ is the CDF of a normal distribution with $\theta$ $(\mu, \sigma^2)$, Equation (11) can be written as:

$$\mathrm{crps}(\mathcal{N}(\mu, \sigma^2), y) = \sigma\left(\frac{y-\mu}{\sigma}\left(2\Phi\left(\frac{y-\mu}{\sigma}\right) - 1\right) + 2\varphi\left(\frac{y-\mu}{\sigma}\right) - \frac{1}{\sqrt{\pi}}\right) \tag{12}$$

where $\varphi$ and $\Phi$ denote the PDF and the CDF, respectively. The CRPS is negatively oriented in a similar manner to the MAE, where a smaller value is better.

The continuous ranked probability skill score (CRPSS) is used to measure the probabilistic skill relative to a reference predictive distribution $F_{ref}$:

$$\mathrm{CRPSS}(F, y) = 1 - \frac{\mathrm{CRPS}(F, y)}{\mathrm{CRPS}\left(F_{ref}, y\right)} \tag{13}$$

The CRPSS is positively oriented, which means that positive values indicate an improvement on the reference forecast. We use the raw ensemble forecasts as the reference forecasts.

Here, we assess the application of EMOS and two machine learning post-processing methods to 1–35 day forecasts of the 2-m maximum air temperature for all land grid points over East Asia (15–60° N, 70–140° E) using the raw ensemble forecast from EMC-GEFS described by Zhu [44]. The calendar year 2016 is used as the validation period. Noted that the EMC-GEFS model initializes once a week so the actual validation period is 53 days.

## 4. Results

### *4.1. Overall Performance of EMOS, the Neural Network and NGBoost*

Table 1 provides the summary measures of the deterministic-style forecast accuracy. For a better visualization and understanding, we average the MAE and RMSE with a lead time of one week instead of one day. The MAE in the first lead week (week 1) is the average MAE of each grid point in the study area during the validation period, with a lead time of 1–7 days. The deterministic neural network forecast clearly gives the best performance, with the mean MAE 24% and 9% reduced than the mean of the raw ensemble and the EMOS forecasts, respectively, for lead times of 1–35 days. The NGBoost method shows similar results, whereas the neural network method performs slightly better for longer lead times.

**Table 1.** Comparison of deterministic forecasts of the 2-m maximum air temperature over East Asia for the year 2016 at different lead times. The values representing the best performance are marked in bold. ENS, raw ensemble; EMOS, ensemble model output statistics; NGB, natural gradient boosting; NN, neural networks; MAE, mean absolute error; RMSE, root-mean-square-error.

| | **Week 1** | | **Week 2** | | **Week 3** | | **Week 4** | | **Week 5** | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **MAE** | **RMSE** | **MAE** | **RMSE** | **MAE** | **RMSE** | **MAE** | **RMSE** | **MAE** | **RMSE** |
| **ENS** | 2.91 | 3.52 | 3.12 | 3.80 | 3.31 | 4.05 | 3.49 | 4.30 | 3.64 | 4.50 |
| **EMOS** | 2.20 | 2.84 | 2.49 | 3.21 | 2.78 | 3.57 | 3.05 | 3.90 | 3.26 | 4.17 |
| **NGB** | 2.05 | **2.63** | 2.30 | 2.93 | 2.54 | 3.24 | 2.77 | 3.53 | 2.97 | 3.77 |
| **NN** | **2.05** | 2.64 | **2.28** | **2.92** | **2.52** | **3.22** | **2.75** | **3.50** | **2.93** | **3.73** |

The comparison of the CRPS between these three post-processing methods and the raw ensemble is presented in Figure 3. It shows that the raw ensemble forecasts are of great uncertainty and all the post-processing methods are able to reduce the CRPS values relative to the raw ensemble for all lead times, especially the two machine learning methods. The improvement in performance could be divided into two parts: weeks 1–2 and weeks 3–5. At a lead time of weeks 1–2, the prediction skill of all the post-processing methods and the raw ensemble decreases with increasing lead time in terms of the CRPS. All the post-processing methods reduced the forecast bias sharply. The neural network and NGBoost methods perform best and increase the available lead time for about 10 days. EMOS achieves similar results to the two machine learning methods in week 1 but performs relatively poorly following week 2. The prediction skill of the post-processing methods and the raw ensemble tends to be stable with little variance as the lead time increases over week 3.

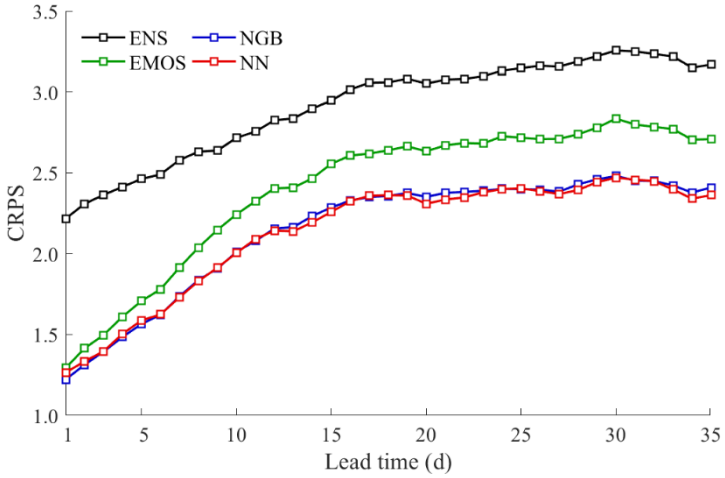


**Figure 3.** Mean continuous ranked probability skill score of the different post-processing methods for the 2-m maximum air temperature at all land grid points over East Asia in 2016 with lead times of 1–35 days. CRPS, continuous ranked probability score; ENS, raw ensemble; EMOS, ensemble model output statistics; NGB, natural gradient boosting; NN, neural networks.

Table 2 presents the CRPS and the coverage and average width of the 88.33% central prediction intervals to assess the accuracy and sharpness, respectively. NGBoost reduces the CRPS score by 39% and 7.5%, respectively, compared with the raw ensemble and the EMOS. The neural network achieves similar results in terms of the CRPS, whereas the NGBoost prediction intervals show better coverage. The raw ensemble prediction intervals are narrow and the accuracy and coverage are unsatisfactory. By contrast, the NGBoost prediction intervals are not much wider than those of the raw ensemble, presenting better CRPS and coverage scores.

**Table 2.** Comparison of probabilistic forecasts of the 2-m maximum air temperature over East Asia for the whole year of 2016 at different lead times. The values representing the best performance are marked in bold. CRPS, continuous ranked probability score; ENS, raw ensemble; EMOS, ensemble model output statistics; NGB, natural gradient boosting; NN, neural network; MAE, mean absolute error; RMSE, root-mean-square-error.

| | **Week 1** | **Week 2** | **Week 3** | **Week 4** | **Week 5** |
|---|---|---|---|---|---|
| | | | ***CRPS*** | | |
| **ENS** | 2.40 | 2.51 | 2.61 | 2.71 | 2.80 |
| **EMOS** | 1.60 | 1.81 | 2.02 | 2.21 | 2.36 |
| **NGB** | **1.48** | **1.65** | **1.82** | 1.98 | 2.12 |
| **NN** | 1.49 | 1.66 | 1.83 | **1.98** | **2.11** |
| | | | ***Coverage at* 88.33% *Prediction Interval*** | | |
| **ENS** | 41.73 | 47.73 | 52.45 | 55.99 | 58.28 |
| **EMOS** | 67.88 | 67.76 | 67.91 | 67.99 | 67.94 |
| **NGB** | **80.30** | **80.06** | **79.98** | **79.94** | **79.71** |
| **NN** | 76.43 | 76.92 | 77.48 | 78.51 | 78.78 |
| | | | ***Average Width at* 88.33% *Prediction Interval*** | | |
| **ENS** | **2.16** | **2.75** | **3.29** | **3.77** | **4.12** |
| **EMOS** | 2.87 | 3.27 | 3.68 | 4.06 | 4.34 |
| **NGB** | 3.52 | 3.94 | 4.38 | 4.78 | 5.10 |
| **NN** | 3.35 | 3.76 | 4.19 | 4.64 | 4.95 |

CRPSS is useful in probabilistic forecasts of multi-category events. Positive values of CRPSS indicate that the forecasts are better than the reference prediction. Figure 4 shows the mean CRPSS of different post-processed forecasts compared with the raw ensemble in different lead weeks. The two machine learning post-processing methods achieve the best and similar results, with mean CRPSS values >0.18 at each lead week. EMOS has a poorer performance, although the mean CRPSS values are >0.08. These results suggest that all the post-processing methods improve the raw ensemble probabilistic forecasts and that the two machine learning methods show the best performance.

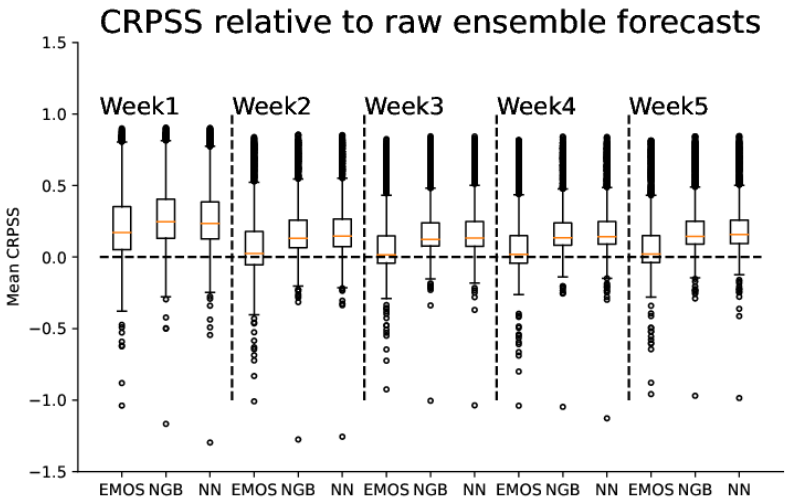


**Figure 4.** Boxplots of the mean continuous ranked probability skill score of all the post-processing models at different lead times using the raw ensemble as the reference forecast. CRPSS, continuous ranked probability skill score; EMOS, ensemble model output statistics; NGB, natural gradient boosting; NN, neural network.

The PIT histograms of the raw and post-processed forecasts used to assess the calibration are shown in Figure 5. The raw ensemble forecasts are under-dispersed, as indicated by the U-shaped verification rank histogram at almost all the selected lead days. This means that the observations often fall outside the range of the raw ensemble. By contrast, the two machine learning post-processed forecast distributions are better calibrated and the corresponding PIT histograms show much smaller deviations from uniformity at each selected lead day. From this perspective, by calibrating the under-dispersive raw ensemble forecasts with nonlinear functions in the neural networks and NGBoost frames, we obtain more idealized ensemble forecasts which reduce the uncertainty and improve the accuracy. The poor performance of EMOS in the PIT histograms with the increasing lead days may reflect the disadvantages of the traditional method that is based on the multi-linear regression techniques.

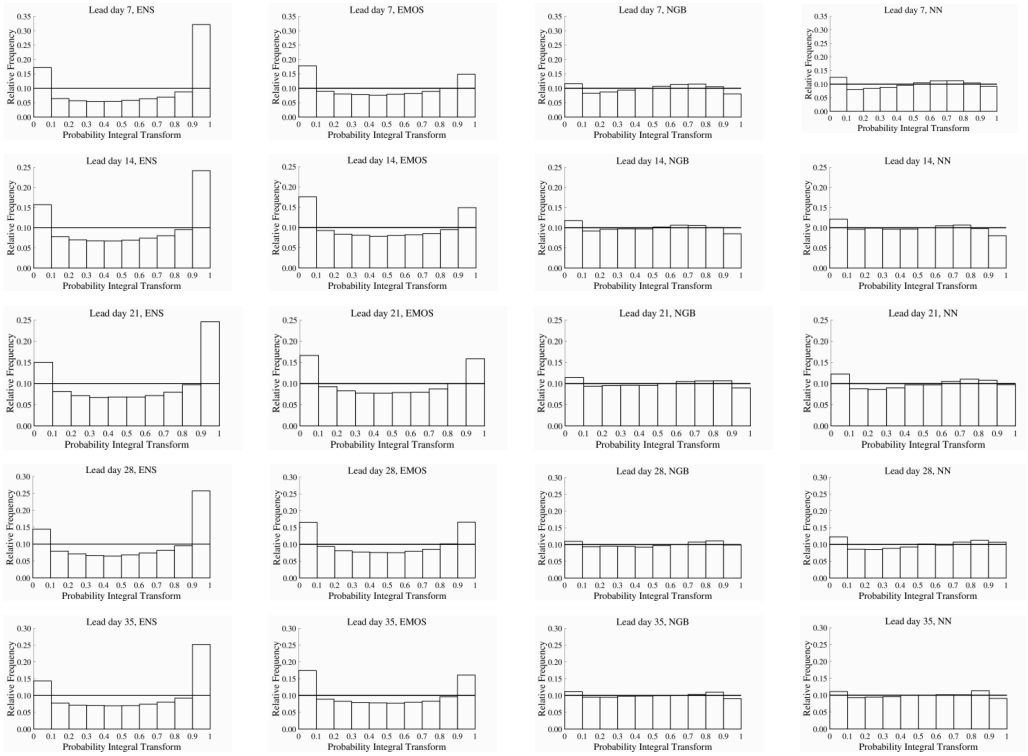


**Figure 5.** Probability integral transform (PIT) histograms of extended range 2-m maximum air temperature probabilistic forecasts for all land grid points over East Asia using different post-processing methods at different lead times. ENS, raw ensemble; EMOS, ensemble model output statistics; NGB, natural gradient boosting; NN, neural networks.

### *4.2. Spatiotemporal Characteristics*

Figure 6 presents the daily CRPS of the raw ensemble and post-processed forecasts in the validation period at different lead days. The method presented by Vigaud et al. (2017) is applied in this study. This evaluates the weekly forecasts initialized in the year 2016, which contains 52 validation periods. Figure 6 shows that the raw ensemble has the poorest performance at each selected lead day. The post-processed forecasts are consistent with the raw ensemble, but perform much better in reducing the CRPS of the raw ensemble, particularly for shorter lead times. The two machine learning methods achieve the best and similar results on most days during the validation periods. The improvements are stable even at long lead times.

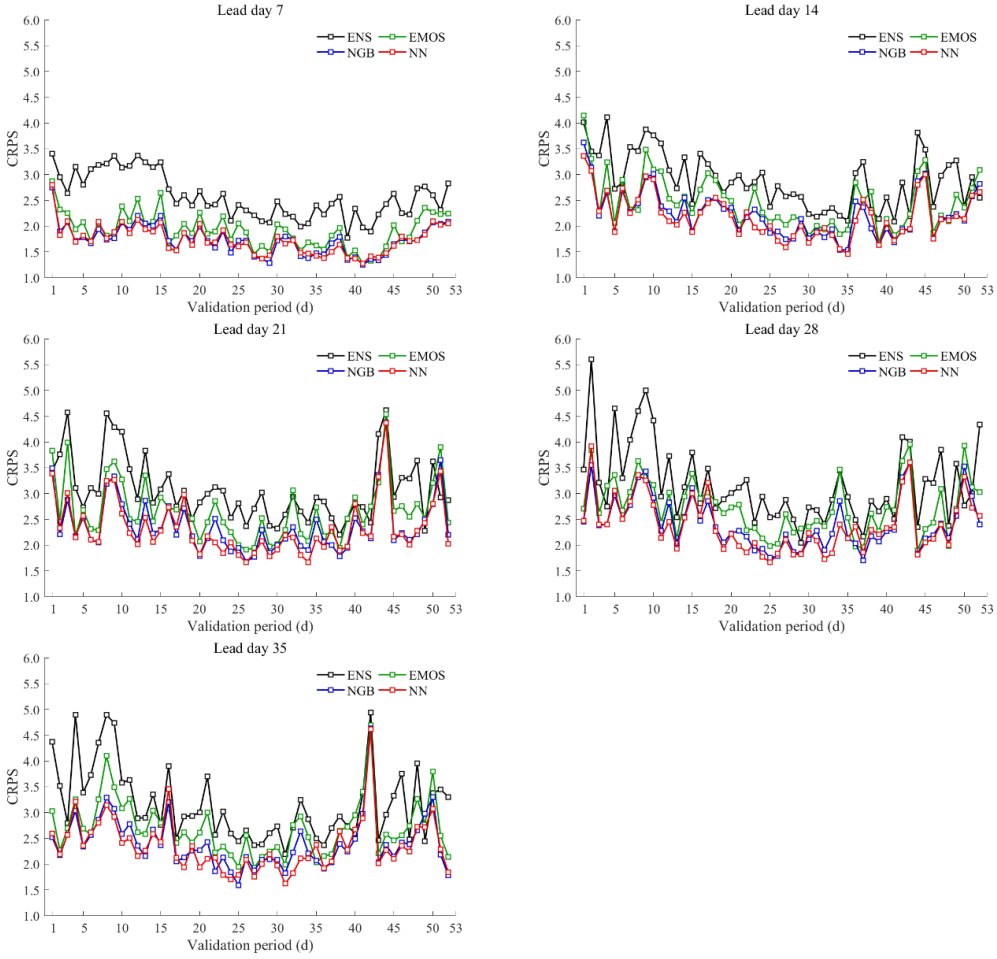


**Figure 6.** Mean continuous ranked probability skill score of the 2-m maximum air temperature probabilistic forecasts for all land grid points over East Asia on each calendar day during the validation period for the raw ensemble and using different post-processing methods at different lead times. CRPS, continuous ranked probability score; ENS, raw ensemble; EMOS, ensemble model output statistics; NGB, natural gradient boosting; NN, neural networks.

Figure 7 shows the spatial distributions of the CRPSS over East Asia for the 2-m maximum air temperature using different post-processing methods, taking the raw ensemble as the reference forecast. Great improvements are achieved by all the post-processing methods in week 1. Most of the grid points in the study area obtain a positive CRPSS, especially those on the Tibetan Plateau. The two machine learning methods reduce the negative CRPSS area more than the EMOS forecast, which indicates that the NGBoost and neural network methods are practicable in broader regions. In week 2 and later, the EMOS forecast skill decreases rapidly and performs poorly in most of Russia, Mongolia and mid-eastern China. However, the two machine learning methods still perform well in those regions. Although the improvements decrease, the post-processed forecast of the two machine learning methods give a positive CRPSS for most grid points in the study area. The neural network forecasts perform better in the northwest of China and most of Russia than the NGBoost outputs. The results indicate that these two machine learning methods could make improvements in complex terrains (such as mountainous areas, deserts) which are often of poor prediction skills. We consider that the improvements describe the nonlinearity to some extent as the thermodynamics and dynamics processes often have.

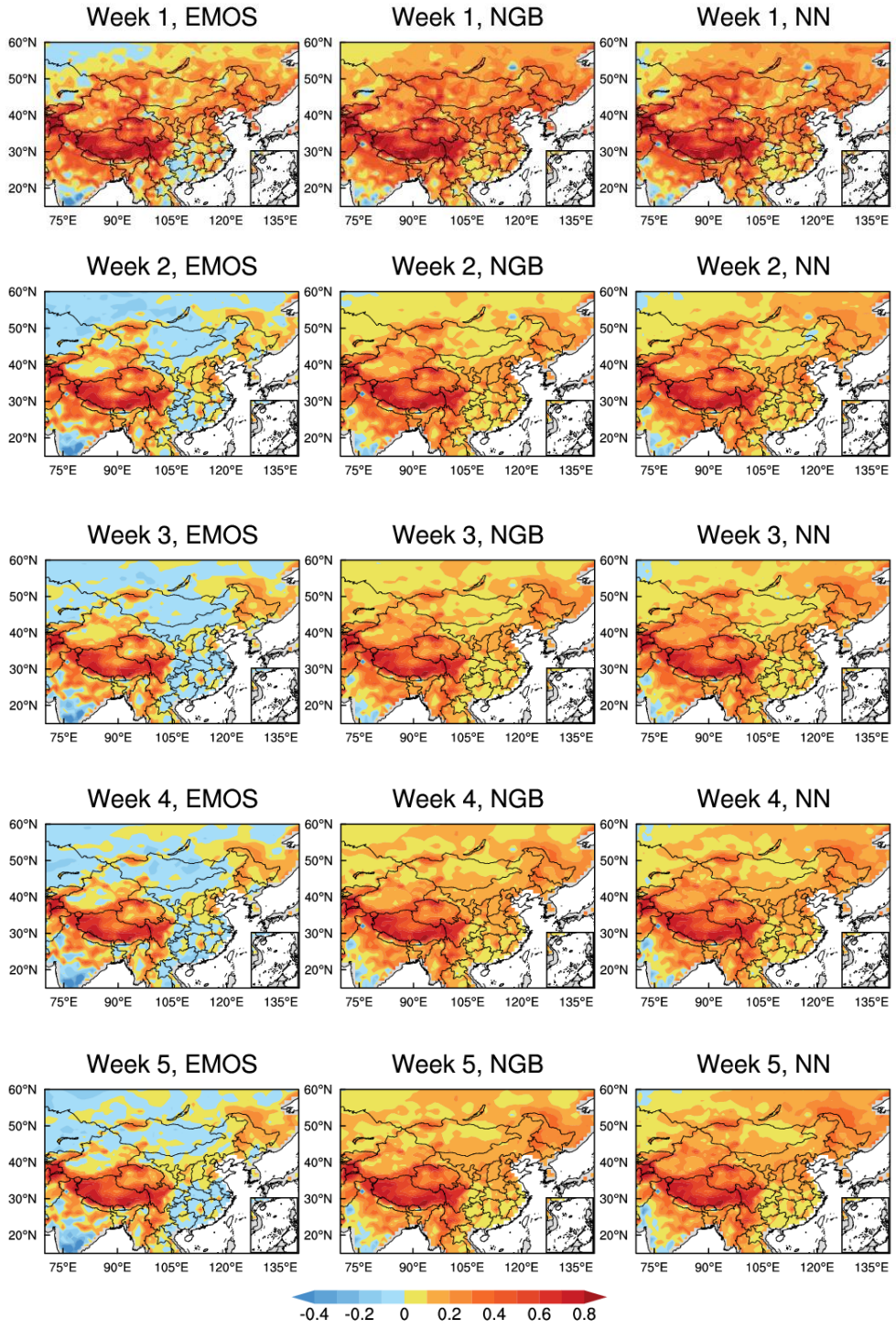


**Figure 7.** Continuous ranked probability skill score distribution of the 2-m maximum air temperature probabilistic forecasts for all land grid points over East Asia verified against the raw ensemble using different post-processing methods at different lead times. EMOS, ensemble model output statistics; NGB, natural gradient boosting; NN, neural networks.

The performance of the neural network is similar to that of NGBoost and thus the best model distribution is introduced to provide a simple comparison. The maps with the best-performing models, in terms of the mean CRPS for each land grid point at different lead weeks, are shown in Figure 8. The two machine learning methods provide the best predictions for most of the grid points. Regardless

of the lead week, the two machine learning methods perform better at >90% of the grid points, with clearer advantages at longer lead times. The NGBoost method performs better in week 1, whereas the neural network method is better in week 2. The two machine learning methods are comparable for longer lead times. For most of the areas, such as the Tibetan Plateau, the Yunnan–Guizhou Plateau, the Loess Plateau, the Inner Mongolia Plateau, the central Siberian Plateau and the East Siberian mountainous area, neural networks behave as the best model despite the overwhelming performance in week 1. There are also some differences over the Tibetan Plateau, for instance, the best model over the west of it tends to be changing from the neural network forecasts to that of NGBoost. In the meantime, over the areas including the northwest of China where the terrains comprising mountains, deserts and basins (in the Xinjiang Province), the neural networks model behaves better, while the NGBoost method dominates in the basins, such as the Qaidam Basin in Qinghai Province and the Sichuan basin in the southeast center of China. For the most populated areas, such as the plains and hilly areas in China, neural networks and NGBoost perform competitively. In all, these two machine learning methods have their own benefits in different areas, but both of them are good at dealing with the complex terrains and adjusting the ensemble spread to provide more reasonable forecasts.

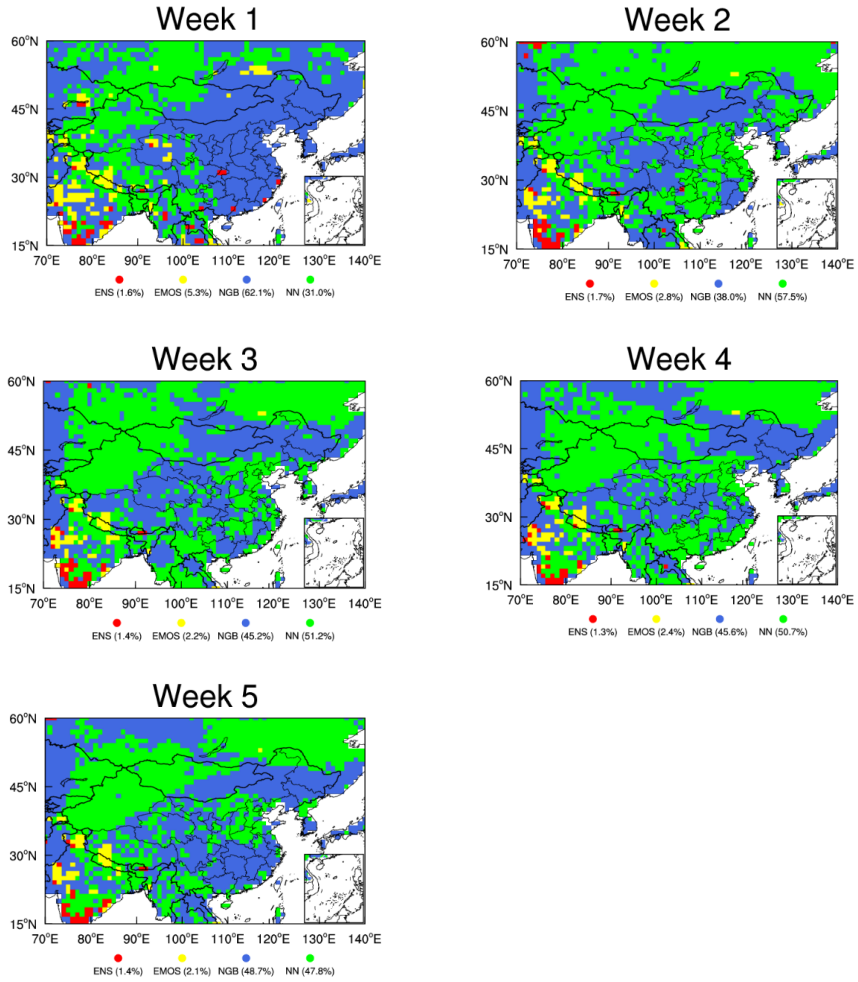


**Figure 8.** Percentage of the best-performing post-processing methods over the land surface in terms of the mean continuous ranked probability score over validation period at different lead times. The percentage of the best-performing forecast of each method is listed under the mosaic image. ENS, raw ensemble; EMOS, ensemble model output statistics; NGB, natural gradient boosting; NN, neural network.

## 5. Conclusions

We compare the performance of the EMOS method and other two methods based on machine learning (a neural network and NGBoost) in extended range temperature probabilistic forecasts over East Asia. The EMOS method has been widely used and is already able to reduce systematic bias and is therefore used as the benchmark for the machine learning methods. Both the neural network and NGBoost improve the forecast skills compared with EMOS over the entire study domain and for each forecast lead time.

For convenience, the lead time is separated into five groups from the first lead week to the fifth lead week and the scoring rules are averaged over the lead time intervals. From the perspective of deterministic forecast, MAE and RMSE are used to assess the accuracy of forecasts from the raw ensembles, EMOS, NGBoost and the neural network. The neural network and NGBoost outperform both the raw ensemble and EMOS outputs and the neural network is slightly better than NGBoost. In terms of convenience and reliability, the neural network is more appropriate and superior for extended range temperature forecasts. However, NGBoost is more flexible and is designed to develop probabilistic forecasts. From the perspective of probabilistic forecasts, both the neural network and NGBoost are superior in terms of CRPS and coverage at the 88.33% prediction interval over the raw ensemble outputs and EMOS results. NGBoost shows advantages over the neural network from the first to third lead weeks in CRPS and the entire forecast's lead weeks in terms of coverage. The raw ensemble forecasts perform better in terms of the under-dispersive spread but give poorer scores. After calibration, EMOS, the neural network and NGBoost expand the spread into a more correct status.

The spatiotemporal distributions are investigated over multiple lead times. The raw ensemble forecasts perform the worst, offering a space for EMOS, the neural network and NGBoost to make progress. The gaps among the three methods decrease with increasing lead times from seven to 28 days. A greater improvement is seen for shorter lead times. The CRPSSs are characterized by spatial variance but are still practical and improved. All the three methods show remarkable forecast skill over the most areas, especially over the Tibetan Plateau. Both the neural network and NGBoost outputs perform better than the EMOS output over East Asia. With increasing forecast lead times, the forecast skill of EMOS decreases remarkably from the first to the second lead week, whereas the other two machine learning methods show sustainably better forecast skills.

The NGBoost and neural network give an outstanding performance in extended range probabilistic forecasts over East Asia for the variable of interest, which can be fitted by a normal distribution. The neural network and NGBoost are well matched. It is difficult to distinguish which one of these two methods is better and we therefore introduced the best model distribution to provide a simple comparison. The neural network and NGB account for almost 90% of the area and they perform better in different lead weeks. However, further investigation of the machine learning applications still required for those variables with a non-normal distribution (e.g., precipitation, wind speed and wind direction).

## 6. Discussion

This paper demonstrates the application of two machine learning methods, the neural networks and NGBoost, to the extended range 2-m maximum air temperature probabilistic forecasts. Following the EMOS frame proposed by Gneiting et al. [11], the neural networks and NGBoost are two parametric methods which directly calibrate the individual raw ensemble members by optimizing a proper scoring rule. However, neural networks and NGBoost show advantages in model robustness and the searching for optimized parameters.

As shown in the flowchart of neural networks (see Figure 1), a neural network consists of two main parts, the forward and backward propagation. For a given forecast time and station, the inputs, the 11 raw ensemble forecasts, are multiplied by $m$ (the node number in the hidden layer) randomly initialized weights $W$ and then activated by a nonlinear function ReLU in the forward propagation to predict the mean $\mu$ and standard deviation $\sigma$. Over the training dataset, the mean CRPS is

computed as the scoring rule by the predictive $\mu$ and $\sigma$ with corresponding observations. In our study, the Adam optimization is applied to gradient descent in the backward propagation for its learning rate varies with training which helps accelerate the convergence.

It is notable that the CRPS is not a common lose function in machine learning. Since the traditional neural networks are incapable for probabilistic forecasting, here we introduce a close form of the CRPS for a Gaussian distribution which helps to extend the neural networks to probabilistic temperature forecast. Furthermore, the inputs of neural networks are more arbitrary, which can add auxiliary predictors to improve prediction skills [24].

This study tackles probabilistic temperature forecasting in atmospheric sciences using NGBoost. It helps make up the gap of gradient boosting method (GBM) in generic probability prediction. The outstanding performance of NGBoost in our study confirms its ability to solve the practical problems. It is a promising machine learning method for probabilistic forecasting for its flexibility and modularization.

Different to neural networks in the forward propagation, NGBoost introduces a weak base learner (for instance, the tree model) to replace the activation function of the neural network frame. By integrating the base learners, we obtain the parameters $\mu$ and $\sigma$ of the predictive PDFs. Another difference lies in the parameter optimization where the natural gradient is introduced in NGBoost to overcome the difficulty of simultaneous boosting the two predictive parameters ($\mu$ and $\sigma$) from the base learners, which is a challenge to the existing GBMs. Furthermore, the natural gradient helps to reflect how the space of distribution moves when updating [25].

For the probabilistic temperature forecast here, the results of bias correction using neural networks and NGBoost are remarkably superior to the traditional EMOS. Neural networks and NGBoost could represent a high nonlinear relationship which is deficiently described in the traditional linear models. Maybe this is the reason why machine learning methods applied in this study perform better. Plus, the machine learning methods are more flexible and the objective function could be adjusted according to the problems to be solved. Compared with the neural networks, the NGBoost is a more integrated and modular method which is especially advantaged with a small training dataset. However, the NGBoost is limited in some skewed probability distribution, for instance, precipitation and wind speed. The neural networks are more suitable for the flexible cases which have more arbitrary predictors and strong nonlinearity.

**Author Contributions:** Conceptualization, X.Z.; Formal analysis, T.P. and Y.J.; Funding acquisition, X.Z.; Investigation, T.P. and Y.J.; Validation, Y.J. and L.J.; Writing—original draft, T.P. and Y.J.; Writing—review and editing, T.P., Y.J. and Y.T. All authors have read and agreed to the published version of the manuscript.

**Funding:** This study was jointly supported by the National Key R&D Program of China (Grant No. 2017YFC1502000) and National Natural Science Foundation of China (Grant No. 41575104).

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

## References

1. DelSole, T.; Trenary, L.; Tippett, M.K.; Pegion, K. Predictability of week-3–4 average temperature and precipitation over the contiguous United States. *J. Clim.* **2017**, *30*, 3499–3512. [CrossRef]
2. Black, J.; Johnson, N.C.; Baxter, S.; Feldstein, S.B.; Harnos, D.S.; L'Heureux, M.L. The predictors and forecast skill of Northern Hemisphere teleconnection patterns for lead times of 3–4 weeks. *Mon. Weather Rev.* **2017**, *145*, 2855–2877. [CrossRef]
3. National Research Council. *Assessment of Intraseasonal to Interannual Climate Prediction and Predictability*; The National Academies Press: Washington, DC, USA, 2010; ISBN 978-0-309-15183-2.
4. Brunet, G.; Shapiro, M.; Hoskins, B.; Moncrieff, M.; Dole, R.; Kiladis, G.N.; Kirtman, B.; Lorenc, A.; Mills, B.; Morss, R.; et al. Collaboration of the Weather and Climate Communities to Advance Subseasonal-to-Seasonal Prediction. *Bull. Amer. Meteorol. Soc.* **2010**, *91*, 1397–1406. [CrossRef]

5. The National Academies of Sciences, Engineering, Medicine. *Next Generation Earth System Prediction: Strategies for Subseasonal to Seasonal Forecasts*; The National Academies Press: Washington, DC, USA, 2016; ISBN 978-0-309-38880-1.
6. Mariotti, A.; Ruti, P.M.; Rixen, M. Progress in subseasonal to seasonal prediction through a joint weather and climate community effort. *npj Clim. Atmos. Sci.* **2018**, *1*, 4. [CrossRef]
7. Pegion, K.; Sardeshmukh, P.D. Prospects for improving subseasonal predictions. *Mon. Weather Rev.* **2011**, *139*, 3648–3666. [CrossRef]
8. Li, S.; Robertson, A.W. Evaluation of Submonthly Precipitation Forecast Skill from Global Ensemble Prediction Systems. *Mon. Weather Rev.* **2015**, *143*, 2871–2889. [CrossRef]
9. Pegion, K.; Kirtman, B.P.; Becker, E.; Collins, D.C.; Lajoie, E.; Burgman, R.; Bell, R.; Delsole, T.; Min, D.; Zhu, Y.; et al. The subseasonal experiment (SUBX). *Bull. Amer. Meteorol. Soc.* **2019**, *100*, 2043–2060. [CrossRef]
10. Raftery, A.E.; Gneiting, T.; Balabdaoui, F.; Polakowski, M. Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Weather Rev.* **2005**, *133*, 1155–1174. [CrossRef]
11. Gneiting, T.; Raftery, A.E.; Westveld, A.H.; Goldman, T. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Weather Rev.* **2005**, *133*, 1098–1118. [CrossRef]
12. Krishnamurti, T.N.; Kishtawal, C.M.; LaRow, T.E.; Bachiochi, D.R.; Zhang, Z.; Williford, C.E.; Gadgil, S.; Surendran, S. Improved weather and seasonal climate forecasts from multimodel superensemble. *Science* **1999**, *285*, 1548–1550. [CrossRef]
13. Zhi, X.; Qi, H.; Bai, Y.; Lin, C. A comparison of three kinds of multimodel ensemble forecast techniques based on the TIGGE data. *Acta Meteorol. Sin.* **2012**, *26*, 41–51. [CrossRef]
14. He, C.; Zhi, X.; You, Q.; Song, B.; Fraedrich, K. Multi-model ensemble forecasts of tropical cyclones in 2010 and 2011 based on the Kalman Filter method. *Meteorol. Atmos. Phys.* **2015**, *127*, 467–479. [CrossRef]
15. Ji, L.; Zhi, X.; Simmer, C.; Zhu, S.; Ji, Y. Multimodel Ensemble Forecasts of Precipitation Based on an Object-Based Diagnostic Evaluation. *Mon. Weather Rev.* **2020**, *148*, 2591–2606. [CrossRef]
16. Ji, L.; Zhi, X.; Zhu, S.; Fraedrich, K. Probabilistic precipitation forecasting over East Asia using Bayesian model averaging. *Weather Forecast.* **2019**, *34*, 377–392. [CrossRef]
17. Neapolitan, R.E.; Neapolitan, R.E. Neural Networks and Deep Learning. *Artif. Intell.* **2018**, 389–411. [CrossRef]
18. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
19. Angermueller, C.; Pärnamaa, T.; Parts, L.; Stegle, O. Deep learning for computational biology. *Mol. Syst. Biol.* **2016**, *12*, 878. [CrossRef]
20. Goh, G.B.; Hodas, N.O.; Vishnu, A. Deep learning for computational chemistry. *J. Comput. Chem.* **2017**, *38*, 1291–1307. [CrossRef]
21. Chao, Z.; Pu, F.; Yin, Y.; Han, B.; Chen, X. Research on real-time local rainfall prediction based on MEMS sensors. *J. Sens.* **2018**, *2018*, 1–9. [CrossRef]
22. Shi, X.; Gao, Z.; Lausen, L.; Wang, H.; Yeung, D.Y.; Wong, W.K.; Woo, W.C. Deep learning for precipitation nowcasting: A benchmark and a new model. In Proceedings of the neural information processing systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 5617–5627.
23. Akbari Asanjan, A.; Yang, T.; Hsu, K.; Sorooshian, S.; Lin, J.; Peng, Q. Short-Term Precipitation Forecast Based on the PERSIANN System and LSTM Recurrent Neural Networks. *J. Geophys. Res. Atmos.* **2018**, *123*, 12543–12563. [CrossRef]
24. Rasp, S.; Lerch, S. Neural networks for postprocessing ensemble weather forecasts. *Mon. Weather Rev.* **2018**, *146*, 3885–3900. [CrossRef]
25. Duan, T.; Avati, A.; Ding, D.Y.; Thai, K.K.; Basu, S.; Ng, A.Y.; Schuler, A. NGBoost: Natural Gradient Boosting for Probabilistic Prediction. Available online: https://arxiv.org/abs/1910.03225 (accessed on 9 June 2020).
26. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
27. Ren, L.; Sun, G.; Wu, J. RoNGBa: A Robustly Optimized Natural Gradient Boosting Training Approach with Leaf Number Clipping. Available online: https://arxiv.org/abs/1912.02338 (accessed on 9 June 2020).
28. Fan, Y.; van den Dool, H. A global monthly land surface air temperature analysis for 1948–present. *J. Geophys. Res. Atmos.* **2008**, *113*, D01103. [CrossRef]
29. 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]

30. Hamill, T.M.; Wilks, D.S. A probabilistic forecast contest and the difficulty in assessing short-range forecast uncertainty. *Weather Forecast.* **1995**, *10*, 620–631. [CrossRef]
31. Stefanova, L.; Krishnamurti, T.N. Interpretation of seasonal climate forecast using Brier skill score, the Florida State University Superensemble, and the AMIP-I dataset. *J. Clim.* **2002**, *15*, 537–544. [CrossRef]
32. Jordan, A.; Krüger, F.; Lerch, S. Evaluating probabilistic forecasts with scoringRules. *J. Stat. Softw.* **2019**, *90*, 1–37. [CrossRef]
33. Kingma, D.P.; Ba, J.L. Adam: A method for stochastic optimization. In Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015), San Diego, CA, USA, 7–9 May 2015; pp. 1–15.
34. Amari, S. Natural Gradient Works Efficiently in Learning. *Neural Comput.* **1998**, *10*, 251–276. [CrossRef]
35. Martens, J. New insights and perspectives on the natural gradient method. Available online: https://arxiv.org/abs/1412.1193 (accessed on 9 June 2020).
36. Gneiting, T.; Raftery, A.E. Strictly Proper Scoring Rules, Prediction, and Estimation. *J. Am. Stat. Assoc.* **2007**, *102*, 359–378. [CrossRef]
37. Gebetsberger, M.; Messner, J.W.; Mayr, G.J.; Zeileis, A. Estimation Methods for Nonhomogeneous Regression Models: Minimum Continuous Ranked Probability Score versus Maximum Likelihood. *Mon. Weather Rev.* **2018**, *146*, 4323–4338. [CrossRef]
38. Gneiting, T.; Balabdaoui, F.; Raftery, A.E. Probabilistic and sharpness forecasts, calibration. *J. R. Stat. Soc. Ser. B-Stat. Methodol.* **2013**, *69*, 243–268. [CrossRef]
39. Anderson, J.L. A Method for Producing and Evaluating Proababilistic Forecasts from ensemble Model Integrations. *J. Clim.* **1996**, *9*, 1518–1530. [CrossRef]
40. Hamill, T.M.; Colucci, S.J. Verification of Eta–RSM Short-Range Ensemble Forecasts. *Mon. Weather Rev.* **1997**, *125*, 1312–1327. [CrossRef]
41. Talagrand, O.; Vautard, R.; Strauss, B. Evaluation of probabilistic prediction systems. In Proceedings of the Workshop on Predictability, Reading, UK, 20–22 October 1997; p. 12555.
42. Hamill, T.M. Interpretation of Rank Histograms for Verifying Ensemble Forecasts. *Mon. Weather Rev.* **2001**, *129*, 550–560. [CrossRef]
43. Hersbach, H. Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems. *Weather Forecast.* **2000**, *15*, 559–570. [CrossRef]
44. Zhu, Y.; Zhou, X.; Li, W.; Hou, D.; Melhauser, C.; Sinsky, E.; Peña, M.; Fu, B.; Guan, H.; Kolczynski, W.; et al. Toward the Improvement of Subseasonal Prediction in the National Centers for Environmental Prediction Global Ensemble Forecast System. *J. Geophys. Res. Atmos.* **2018**, *123*, 6732–6745. [CrossRef]