Transformer for Sub-Seasonal Extreme High Temperature Probabilistic Forecasting Over Eastern China

Wei Jin

Fujian Normal University

Wei Zhang

Fujian Key Laboratory of Severe Weather

Jie Hu

Fujian Normal University

Jiazhen Chen

Fujian Normal University

Bin Weng

Fujian Normal University

Jianyun Gao

Fujian Key Laboratory of Severe Weather

Tianqiang Huang (✉ fjhtq@fjnu.edu.cn)

Fujian Normal University

Research Article

Keywords: Deep learning, Sub-season, Transformer, Extreme high temperature

Posted Date: January 10th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1230879/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Transformer for Sub-Seasonal Extreme High Temperature Probabilistic Forecasting Over Eastern China

Wei Jin^{1†}, Wei Zhang^{2†}, Jie Hu¹, Jiazhen Chen¹, Bin Weng¹, Jianyun Gao²
and Tianqiang Huang^{1*}

¹Fujian Normal University, Fuzhou, 350007, Fujian, China.

²Fujian Key Laboratory of Severe Weather, Fuzhou, 350007, Fujian, China.

*Corresponding author(s). E-mail(s): fjhtq@fjnu.edu.cn;

Contributing authors: jwillking@qq.com; si_su@163.com; hjgujin@gmail.com;
jiazhen_chen@fjnu.edu.cn; binweng@fjnu.edu.cn; fzgaojyun@163.com;

[†]These authors contributed equally to this work.

Abstract

Sub-seasonal high temperature forecasting is significant for early warning of extreme heat weather. Currently, deep learning methods, especially Transformer, have been successfully applied to the meteorological field. Relying on the excellent global feature extraction capability in natural language processing, Transformer may be useful to improve the ability in extended periods. To explore this, we introduce the Transformer and propose a Transformer-based model, called Transformer to High Temperature (T2T). In the details of the model, we successively discuss the use of the Transformer and the position encoding in T2T to continuously optimize the model structure in an experimental manner. In the dataset, the multi-version data fusion method is proposed to further improve the prediction of the model with reasonable expansion of the dataset. The performance of well-designed model (T2T) is verified against the European Centre for Medium-Range Weather Forecasts (ECMWF) and Multi-Layer Perceptron (MLP) at each grid of the 100.5°E to 138°E, 21°N to 54°N domain for the April to October of 2016-2019. For case study initiated from 2 June 2018, the results indicated that T2T is significantly better than ECMWF and MLP, with smaller absolute error and more reliable probabilistic forecast for the extreme high event happened during the third week. Over all, the deterministic forecast of T2T is superior to MLP and ECMWF due to ability of utilize spatial information of grids. T2T also provided a better calibrated probability of high temperature and a sharper prediction probability density function than MLP and ECMWF. All in all, T2T can meet the operational requirements for extended period forecasting of extreme high temperature. Furthermore, our research can provide experience on the development of deep learning in this field and achieve the continuous progress of seamless forecasting systems.

Keywords: Deep learning, Sub-season, Transformer, Extreme high temperature

1 Introduction

Extreme high temperature events have major impacts on human life ([Zhu et al. 2021b](#)). In May

2010, a persistent heatwave swept across north-west India, with highs approaching 50°C, resulting in a rare drought threatening 100 million people. In May 2014, temperatures in many parts

of China broke historical maximum temperatures, with major impacts on agriculture and fisheries. In the summer of 2013, eastern China suffered the worst high temperature in 113 years (Xia et al. 2016). There is an urgent need to find a more effective way to predict the onset of extreme high temperatures in advance, thereby reducing economic and personnel losses.

However, existing researches is dominated by short-range forecasting (Tran et al. 2021; Johnstone and Sulungu 2021), with forecasts of a few hours to a few days meaning that the time left for people to respond is very limited and does not meet society's need for longer lead times. The emergence of weather forecasts that offer longer lead times has become a growing trend, such as sub-seasonal forecasts, which fall between medium-term and seasonal forecasts (Vigaud et al. 2019; Pegion et al. 2019; Vijverberg et al. 2020). As a key component of a seamless forecasting system, sub-seasonal forecasts can provide weeks of advance warning of extreme events.

Probabilistic forecast is essential for solving uncertainty of extended-range forecast. Due to a chaotic characteristic of atmosphere and uncertainties associated with initial conditions and models (Lorenz 1963 1969; Thompson 1957; Smagorinsky 1969), it is difficult for deterministic forecast to reproduce the future atmospheric reality. Besides, the deterministic forecast often carry a high risk, and even cause irreversible losses (Möller and Groß 2016; Boi 2004). Probabilistic forecast can be used as an auxiliary forecasting tool to provide more information for users' decision making, and to a certain extent, it can effectively distinguish the forecastability of the forecast object (Zhu et al. 2002). In recent years, sub-seasonal high temperature probabilistic forecasting methods are mainly based on multi-model integration (Zhu et al. 2018; Ji et al. 2020; Zhu et al. 2021a; Rasp and Lerch 2018). However, these models are difficult to develop and maintain, while the accuracy of forecasting still has much room for improvement (Garmong et al. 2020).

Nowadays, deep learning has been gradually applied to the field of sub-seasonal high temperature forecasting (He et al. 2020; Peng et al. 2020). Among them, Peng et al. (2020) used multilayer perceptron (MLP) to investigate probabilistic forecasting. The method improves the forecasting effectiveness of the model while reducing

the modeling cost, but it only considers the patterns of individual grid, thus ignoring the global correlations between grids. In the field of remote sensing, Transformser achieves outstanding results for its excellent ability to extract global correlations (Ye et al. 2021). In this paper, we introduce the Transformer and propose a new model called Transformer to High Temperature (T2T) for sub-seasonal high temperature probabilistic forecasting.

Section 2 introduces the dataset and the validation methods. The proposed model, framework, and other details are described in Section 3. Section 4 shows the process of optimization and justification of the T2T model, including position encoding, Transformer and multi-version integration of datasets. In Section 5, we demonstrate the superior performance of T2T from individual case analysis to the overall analysis of deterministic and probabilistic forecasts. Lastly, Section 6 provides the main contributions and conclusion of this paper.

2 Data and evaluation methods

2.1 Downloading and processing of data

The study area of this paper is eastern China (21°N-54°N, 100.5°E-138°E), and the study months are from April to October. Two sets of data are used. The first set is six versions of six-hourly 2-m temperature reforecasts downloaded from the European Centre for Medium-Range Weather Forecasts (ECMWF) from the 2015 version to the 2020 version. Each version contains historical hindcast data for the past two decades at a resolution of 1.5°*1.5° (<https://apps.ecmwf.int/datasets>). The second set is the observed daily maximum temperatures of 2248 stations in China provided from the Chinese Meteorological Information Center, bilinearly interpolated to observations with a resolution of 1.5°*1.5°.

The first data set contains one control forecast and ten perturbed forecasts, which we combine to form a forecast with eleven forecast members. After adjusting for time differences between two data sets, we first integrate the six-hourly data

Table 1 Division of the data set

Versions	Data sets divided by year		
	Training sets	Test sets	Invalid sets
2015	1995-2014	-	-
2016	1996-2015	-	-
2017	1997-2015	-	2016
2018	1998-2015	-	2016-2017
2019	1999-2015	-	2016-2018
2020	2000-2015	2016-2019	-

into a daily data set, then generate four weeks of data in seven consecutive-day increments, and finally the eleven forecasts are averaged to obtain the ensemble mean is noted by x as the data to be corrected and also as the forecast result of ECMWF. The second data set generates corresponding weekly observations based on the dates of the first data set as labels for model training.

For the training and testing of the network, we choose all versions from 1995 to 2015 for the training set and the 2020 version from 2016 to 2019 for the test set. Some data is discarded, mainly considering that there can be no crossover of years between the training set and the test set, so as to avoid leakage of data in the test set. The detailed division of the data set is shown in Table 1.

2.2 Determination of extreme high temperature thresholds

The selection of extreme high temperature thresholds at sub-seasonal scales is important. Considering that there were various gaps in the variability of high temperatures at each grid point, we constructed weekly observations by sliding averages every seven days from April to October in the training year. Then we used the 90% percentile of each grid as the extreme high temperature thresholds $THR \in R^{H_0 \times W_0}$ (Wulff and Domeisen 2019), as shown in Fig. 1. Our thresholds are consistent with the high temperatures in East China from July to August 2013, indicating that settings of the thresholds are reasonable (Xia et al. 2016).

2.3 Verification methods

The evaluations in this paper are divided into deterministic and probabilistic evaluations, which correspond to different indicators. Mean absolute error (MAE), root mean square error (RMSE),

Fig. 1 Spatial distribution of extreme high temperature thresholds.

equitable threat score (ETS), anomaly correlation coefficient (ACC), and time horizon correlation coefficient (TCC) were used as indicators for deterministic forecasts, which were assessed in terms of numerical error, fall zone, and correlation; brier score (BS), brier skill score (BSS) and reliability score as indicators for probabilistic forecasts from the perspective of probability error, probabilistic forecast skill, errors and reliability in probability distributions (Rasp and Lerch 2018; Boi 2004; Hamill and Juras 2006; Ji et al. 2019; Specq et al. 2020; WilksDS 2011).

3 T2T model and framework

3.1 Model structure

We are inspired by the Vision Transformer (ViT) (Dosovitskiy et al. 2020) and introduce the Transformer into our model, proposing a model structure with the Transformer as the main body, called Transformer to High Temperature (T2T). The model structure is shown in Fig. 2. The model consists of a layer of convolution module and several Transformer modules. The convolution layer is used as a filter for multidimensional data, fully extracting the information contained in the data itself by sharing weights, feature sampling and generating a new tensor of the required size while

Fig. 2 Diagram of the structure of the T2T model. L is the number of Transformer modules. ω is the weight of the position embedding. The structure consists of a convolutional layer, patch division, semantic embedding, position embedding, Transformer and full connection. In deterministic forecasting, the input is the original ECMWF forecast data and the output is the revised result ; in probabilistic forecasting, the input is the revised result in deterministic forecasting, which is also the expectation of the distribution, and the output is the standard deviation under the distribution.

ensuring a degree of translational invariance to meet the input data size requirements of subsequent modules. Transformer uses a multi-headed attention mechanism to extract features in different feature spaces. It determines different weights for each feature through its relationship to the target data, so it works well in global feature extraction (Ye et al. 2021). The complete calculation process of the model in this paper is as follows equation 1.

$$\begin{aligned}
 a_0 &= [X_{\text{pre}}, PD(\text{Conv}(x))] + \omega X_{\text{pos}}, \omega \in [0, 1] \\
 a_l &= \text{MHA}(\text{LN}(a_{l-1})) + a_{l-1}, l = 1, \dots, L \\
 a_l &= \text{MLP}(\text{LN}(a_l)) + a_l, l = 1, \dots, L \\
 y &= \text{RS}(\text{FC}(a_L))
 \end{aligned} \quad (1)$$

Firstly, our input is the previously generated $x \in R^{1 \times H_0 \times W_0}$. After one convolution operation (Convolution, Conv), the feature map $f_1 \in R^{C_1 \times H_1 \times W_1}$ is obtained. The feature map is then divided (Patch Division, PD) into N smaller feature maps of the same size $f_2 \in R^{C_1 \times P \times P}$ and flattened to one-dimensional vectors. Finally, the semantic embedding $f_3 \in R^{N \times C_2}$ is obtained by using a trainable linear mapping for f_2 . It should be noted that we have added a learnable field at the top of the semantic embedding $X_{\text{pre}} \in R^{1 \times C_2}$.

We have also used position embedding denoted by $X_{\text{pos}} \in R^{(N+1) \times C_2}$, which helps us to record the relative positions of the chunked feature tensor (Dosovitskiy et al. 2020) and add it to the semantic embedding as the input of Transformer. Here, we set weight ω for the position embedding and used it to explore the effect of position embedding on the effects in our research task. In the Transformer section we use the Transformer module in ViT, which consists of the multi-head attention (MHA), the multilayer perceptron (MLP) and the layer normalization (LN). After multiple encodings, our forecast result is output in the form of a tensor $y \in R^{1 \times H_0 \times W_0}$ through fully connected layers (FC) and tensor restructuring (RS).

3.2 Process framework

In deterministic forecasting, x undergoes a deterministic forecast model $T2T_C$ to obtain a corrected forecast result y . In probabilistic forecasting, high temperature variations conform to the laws of variation of the Gaussian distribution (Gneiting et al. 2005; Rasp and Lerch 2018). To establish the association from deterministic to probabilistic forecasting, we enter y as the expectation μ of a Gaussian distribution into the probabilistic forecasting model $T2T_P$, and generate the standard deviation σ under a point-by-point

Fig. 3 Histograms for the evaluation of multiple learning methods and multi-version fusion method. MAE, RMSE, ETS and ACC are the four indicators.

Gaussian distribution from the network under the action of a specific loss function. Ultimately, after obtaining a Gaussian distribution function based on the standard deviation σ and the expectation μ , we integrate for cases greater than the high temperature threshold THR to obtain the grid point-by-grid high temperature probability F . The whole process is expressed in the following equation 2.

$$y = T2T_C(x)$$

$$\mu = y$$

$$\sigma = T2T_P(\mu)$$

$$F(THR) = P(X > THR) = \frac{1}{\sqrt{2\pi}\sigma} \int_{THR}^{+\infty} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt \quad (2)$$

3.3 Other details

In this paper we use MAE as the loss function in deterministic forecasting, and in probabilistic forecasting we use CRPS based on Gaussian distribution as the loss function. The loss values of the forecast results for M results with n -grid points are shown in equation 3, 4. For all experiments we use Adam as the optimizer, 0.0001 as the learning rate and 500 epochs of iterations. In addition, in the model, the number of Transformer modules (L), number of heads of attention mechanism are set to 2, 8, respectively. The main experimental section has two models whose results are compared with ensemble mean of the ECMWF: multilayer perceptron (MLP) and Transformer to High Temperature (T2T), where MLP directly uses the network structure of the reference (Peng et al. 2020).

Fig. 4 Spatial distribution of absolute error for different forecasts initiated from 2 June 2018. The rows indicate the three forecast methods: ECMWF(a, b, c, d), MLP(e, f, g, h) and T2T(i, j, k, l), and the columns indicate the four forecast weeks.

$$\text{Loss} = \frac{1}{MN} \sum_{i=1}^M \sum_{g=1}^N |X_{i,g} - Y_{i,g}| \quad (3)$$

In deterministic forecasting, x is the forecast value and Y is the observed value (Ji et al. 2019).

$$m = \frac{Y - \mu}{\sigma}, t = (N(\mu, \sigma^2), Y)$$

$$\text{CRPS}(t) = \sigma \left(m(2\Phi(m) - 1) + 2\varphi(m) - \frac{1}{\sqrt{\pi}} \right)$$

$$\text{Loss} = \frac{1}{MN} \sum_{i=1}^M \sum_{g=1}^N \text{CRPS}(N_{i,g}(\mu, \sigma^2), Y_{i,g}) \quad (4)$$

In probabilistic forecasting, Y is the observed value, Φ and φ are the predictive cumulative distribution function (CDF) and the probability density function (PDF) of a normal distribution with $N(\mu, \sigma^2)$ (Hamill and Juras 2006).

Table 2 The influence of Transformer. Note T2T (T2T-NT) means the model with (without) Transformer.

Indicators	Models	Week1	Week2	Week3	Week4
MAE	T2T-NT	0.95	1.60	1.89	1.91
	T2T	0.95	1.60	1.87	1.91
RMSE	T2T-NT	1.21	2.04	2.40	2.44
	T2T	1.21	2.04	2.38	2.43
ETS	T2T-NT	0.54	0.28	0.12	0.09
	T2T	0.55	0.30	0.15	0.14
ACC	T2T-NT	0.96	0.90	0.88	0.87
	T2T	0.96	0.90	0.88	0.88
BS	T2T-NT	0.05	0.08	0.09	0.09
	T2T	0.04	0.08	0.09	0.09

4 Experiments for optimizing T2T

4.1 The influence of model's details

In the model, we discuss whether the existence of Transformer and position embedding in T2T is necessary.

T2T is a model with Transformer as the main body. In order to verify whether the existence of

Fig. 5 Spatial distribution of different probabilistic forecast initiated from 2 June 2018. The first row indicates the observations and the red (100%) indicates that the area is a high temperature region. The other rows indicate the three forecast methods: ECMWF(a, b, c, d), MLP(e, f, g, h) and T2T(i, j, k, l). The columns indicate the four forecast weeks.

Transformer plays a positive role in the model, we remove the Transformer in T2T to form a new model, denoted as T2T-NT, and compare the experimental effects of the T2T and T2T-NT. The results are shown in Table 2.

It can be seen that T2T with Transformer performs better than T2T-NT. Especially in the ETS, with the longer extension period, T2T still forecasts better high-temperature fallout areas. Therefore, the Transformer with global feature extraction ability plays an positive role in the

model, which makes the forecast results more meaningful for the extended period.

In Equation 1, we add a weight ω to the use of position embedding, where we set the weights ω to 0, 0.5, and 1 to explore the effect of position embedding on the overall model forecasts under this experimental task. The results are shown in Table 3. The best results are obtained when the weight is 0, and the higher the weight, the worse the indicator results, which indicates that the existence of position embedding in our task does not

Fig. 6 Histogram of deterministic forecast evaluation. MAE, RMSE, ETS, ACC are the four indicators.

Table 3 Results of indicators with different weights ω . MAE, RMSE, ETS, ACC are deterministic forecast indicators, BS is probabilistic forecast indicator.

	ω	Week 1	Week 2	Week 3	Week 4
MAE	0	0.95	1.60	1.87	1.91
	0.5	0.98	1.59	1.89	1.93
	1	1.01	1.62	1.91	1.93
RMSE	0	1.21	2.04	2.38	2.43
	0.5	1.25	2.03	2.40	2.47
	1	1.29	2.08	2.43	2.47
ETS	0	0.55	0.30	0.15	0.14
	0.5	0.53	0.27	0.10	0.08
	1	0.52	0.29	0.12	0.08
ACC	0	0.96	0.90	0.88	0.88
	0.5	0.96	0.90	0.88	0.88
	1	0.96	0.90	0.88	0.88
BS	0	0.04	0.08	0.09	0.09
	0.5	0.05	0.08	0.09	0.09
	1	0.05	0.08	0.09	0.09

effectively improve the forecast performance of the model, but rather has a greater negative impact on the forecast results.

In summary, when ω is 0, T2T without position encoding can be the most outstanding model.

In the field of image classification, position embedding is used to preserve the spatial position relationships between image blocks, thus reducing the additional cost of acquiring position information through network training (Dosovitskiy et al. 2020). Following previous ideas, this paper also uses position embedding to improve forecasting by memorizing the image block position relationships, but in practical experiments we find that the presence of position embedding degrades the model performance. We consider that in the classification problem, the model only needs to identify the target class by mining the key features, so it does not require high clarity of the image and has a high error tolerance. The use of position embedding, which sacrifices some of the sharpness for the positional information of the image block, achieves the same or even better classification results with a high error tolerance for the classification problem.

However, in the field of meteorological forecasting, where there are various numerical relationships underlying the data, it is unwise to

Fig. 7 Spatial distribution of TCC for deterministic forecasting. The columns represent the four forecast weeks. The rows represent the TCC distributions of the three forecast outcomes we compared: ECMWF(a, b, c, d), MLP(e, f, g, h) and T2T(i, j, k, l).

Fig. 8 Histogram of probabilistic forecast evaluation. BS is indicator.

Fig. 9 Spatial distribution of BSS for probabilistic forecasting. The columns represent the four forecast weeks. The rows represent the BSS distributions of the three forecast outcomes we compared: ECMWF(a, b, c, d), MLP(e, f, g, h) and T2T(i, j, k, l). Red represents positive skill and blue represents negative skill.

sacrifice the accuracy of the data by using position embedding. This is because the addition of position embedding to weather data that is already subject to systematic error is tantamount to increasing the systematic error, resulting in a disruption of the original numerical relationships and feature learning, and ultimately a reduction in model performance. Position embedding is therefore not necessary in this task.

4.2 Effectiveness of multi-version integration

To verify the validity of the method, we designed three learning modes based on testing the 2020 version: learning with the 2020 version (same version learning), learning with the 2019 version (cross-version learning), and learning with the 2019 version and then a small amount of re-learning on the 2020 version (transfer learning).

As shown in Fig. 3, we compare the multi-version fusion method of this paper with the three learning methods. For the three newly proposed learning methods, the differences between the individual indicators are small, with the different learning methods showing almost equal levels of forecasting performance with the same test set. This indicates that there is some correlation and portability between versions. In contrast, multi-version fusion learning method expands the dataset sensibly and efficiently by using correlations between versions, resulting in optimal experimental results, especially in the ETS.

5 Performance of T2T

5.1 Case analysis

To analyze the forecasts for deterministic and probabilistic forecasts, we chose a four-week forecast initiated from 2 June 2018 and exhibit the spatial distribution of errors, with the domain of China experiencing a serious serious extreme high event during this forecasting period.

In the analysis of the deterministic forecasts, we compare the absolute errors of the forecasts and observations, as shown in Fig. 4. Compared with the absolute error of ECMWF, the errors of MLP are significantly reduced and lighter in many places, which indicates that the multilayer perceptron works well in the revision problem of multigrid regression and has largely been able to effectively revise the ECMWF forecast data. However, compared with T2T methods, there is still a gap. It can be seen that T2T has the smallest error coverage area and error values.

In the analysis of the probabilistic forecasts, we compare the probabilistic forecasts of the three methods with the observations, as shown in Fig. 5. The first row in red marks the real high temperature areas for four consecutive weeks. In the ECMWF forecasts, the first two weeks are good, with high probability values above 95% for some of the high temperature areas, but the second two weeks are poor, with little correlation with the true high temperature areas even at the maximum probability values. In the MLP forecasts, although the MLP shows its good revisions in the deterministic forecasts, it falls short of expectations in the probabilistic forecasts: the whole maps are covered by low probability values, with the highest probability of only 60% in the first two weeks and a weak correlation with observations at the 50% highest probability in the second two weeks. This shows that the forecast results of the MLP in probability are very poor and even inferior to the original probabilistic forecasts. Compared to the two probabilistic results above, the T2T forecasting methods performs well. Larger high temperature fallout areas and high probability values above 95% are forecast in the first two weeks. In the third week, when other methods are no longer able to predict the area of high temperature based on probabilistic forecasts, the T2T method can still estimate the approximate area of

high temperature based on the maximum probability forecast value of 60% from the features of the area of high temperature learned by the model. Unfortunately, the fourth week forecast results, like several other methods, have little useable value. The current probabilistic forecasts are less satisfactory in the latter two weeks (Guan et al. 2019) in which case the global feature extraction capability of the T2T method has led to a continued trend towards extending the period for probabilistic forecasts with realistic relevance.

Overall, the T2T model achieves the best results for both deterministic and probabilistic forecasts at this initial date, demonstrating its superiority over the other models, which proves that our proposed new model is indeed better suited to the forecasting task of this paper.

5.2 Deterministic forecasting

In Fig. 6, we compare the results of the three methods. The MLP outperforms the ECMWF in all metrics, indicating that the multilayer perceptron is effective in correcting the original forecasts. However, T2T obtains an overall leading result, relying on its stronger global feature extraction capability.

To compare the forecasting effects more visually, in Fig. 7 we show the spatial distribution of TCC for the three methods. Comparing the ECMWF with the MLP, we find that the former is slightly more effective than the latter in the latter two weeks, while in other cases the TCC distributions for both are very similar. We then compare their results with T2T and notice that there is a significant improvement in TCC in the same forecast weeks from ECMWF and MLP to T2T, such as the middle and lower reaches of the Yangtze River and the northeast region.

In contrast to MLP, which underperforms in deterministic forecasting due to the lack of ability to utilize spatial information of grids, T2T makes up for this deficiency and shows excellent error correction ability.

5.3 Probabilistic forecasting

In probabilistic forecasting, we regard the output of a deterministic forecast as the input to a probabilistic model and the expectation of a Gaussian distribution. Then using the model we obtain the standard deviation under that distribution.

Fig. 10 The reliability diagram for high temperature probability forecasts. The reliability comparisons of ECMWF(green), MLP(blue) and T2T(red) are shown respectively. The line graph represents the relationship between the observed relative frequency and the forecast probability, where the black diagonal line is the perfect reliability line. The bar chart indicates the probability distribution of the forecast values. With the frequency of the predicted values as the weight, the scores in the lower right corner of the line graph are the area between methods' curves and the perfect reliability line meaning the difference from the perfect reliability line.

Ultimately, the probabilities are calculated from the distributions. Since our study focus on the dichotomous extreme event and CRPS measures the overall probabilistic performance, so it is not employed here.

Fig. 8 shows the comparison of BS score between ECMWF, MLP and T2T. It is indicated that T2T has much calibrated probability of high temperature for the first week forecast. This revision capability decreases as the forecasting week increases. Nonetheless, T2T perform best for all forecasting weeks.

We engage BSS to evaluate the forecasting skill relative to the climatological probabilities for ECMWF, MLP and T2T. The result is shown in Fig. 9. The ECMWF has a clear distribution of forecast skill, with positive skill dominated by northern China, Jiangsu Province and Anhui Province. MLP has a significant weakening effect on the negative skill of the ECMWF, but inevitably the negative optimization of the northern region results. The forecasting skill of the T2T is significantly improved compared to the ECMWF and MLP: the full positive skill is basically achieved in the first week, and the positive skill is enhanced in weeks 2-4 for northern China and the middle and lower Yangtze River region. Looking again at Fig. 9, we find that a grid point located near the Taihang Mountains consistently exhibits negative skill, indicating the difficulty of probabilistic forecasting at this location. We speculate that this may be due to special circumstances such as local geographical factors.

Reliability assessment is essential for any decision based on forecasting (Weisheimer and Palmer 2014). In addition to the skill of forecast, we analyze the reliability of forecast as in Fig. 10. We set the probabilities in 12 bins: 0.0, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0. For each given bin of forecast probability, reliability is the correspondence between the forecast and the observed high temperature. Lines with different colors in Fig. 10 represent different forecasting reliability. Lines above the diagonal line exhibit under-forecasting bias while below diagonal line exhibit over-forecasting bias. For ECMWF, the part with small probability shows an under-forecasting bias and the part with large probability shows an over-forecasting bias. MLP is dominated by over-forecasting bias, with a few under-forecasting bias.

The reliability of T2T (red line in Fig. 10) is closest to the diagonal line, showing more reliable than ECMWF and MLP. Then, for the histogram of probability distribution, MLP is scattered, mostly concentrated in the medium probability, which hardly provides a clear guidance for decision making. In contrast, ECMWF and T2T are clustered together, with ECMWF in particular having the highest sharpness at small probabilities. Finally, based on the overall scores, we compare the overall reliability of the three forecasting methods. MLP has the largest scores and largest difference from the diagonal and the least reliable forecast results, indicating it is inferior to ECMWF. The scores of T2T are the smallest, suggesting closer to the perfect forecast curve with the best reliability. All in all, we demonstrate the superiority of T2T in probabilistic forecasting from a reliability perspective, giving us more confidence in the new approach.

6 Conclusion

Accurate sub-seasonal high temperature forecasting can improve human ability to cope with extreme high temperature. To address this goal, we explore the application of the Transformer in this field and demonstrate that the Transformer is feasible. In addition, some experimental details that are beneficial to improve the performance of the model are presented, providing experiences and methods for future research. The main contributions are as follows.

1. We propose a multi-version fusion method for datasets and build a week-by-week dataset based on hindcast data from ECMWF and observations from Chinese stations.
2. We pioneered the introduction of Transformer into the field of sub-seasonal forecasting by proposing the T2T model and established a high temperature deterministic and probabilistic forecasting system. The remarkable experimental results show that the new model is indeed competent to forecast extreme high temperatures and our method makes the extended period forecast more meaningful.
3. We experimentally demonstrate that positional embedding, which can improve model performance in the classification domain, does not

achieve the expected results in the multi-grid regression problem of weather forecasting.

In the future, we will build more efficient models and introduce factors such as circulation and topography.

Declarations

- Funding

This study was supported by National Key R&D Program Special Fund Grant (No. 2018YFC1505805), National Natural Science Foundation of China (No. 62072106, No. 61070062), General Project of Natural Science Foundation in Fujian Province (No. 2020J01168), Open project of Fujian Key Laboratory of Severe Weather (No. 2020KFKT04), Natural Science Foundation of Fujian Province of China (No. 2021J01454), Guiding Project on Social Development of Fujian Province (No. 2021Y0057).

- Conflict of interest

The authors declare no competing of interests.

- Ethics approval, Consent to participate, Consent for publication

Not applicable

- Availability of data

ECMWF: <https://apps.ecmwf.int/datasets>

Other related content can be obtained by sending an email to the authors.

- Code availability

The code can be obtained by sending an email to the authors.

- Authors' contributions

WJ and WZ did the experiments, writing, and analysis together, and the other authors helped with touch-ups and other content refinements.

References

- Boi P (2004) Probabilistic temperature forecast by using ground station measurements and ecmwf ensemble prediction system. *Meteorological Applications* 11(4):301–309. <https://doi.org/10.1017/S1350482704001380>
- Dosovitskiy A, Beyer L, Kolesnikov A, et al. (2020) An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* URL <https://arxiv.org/abs/2010.11929>

preprint arXiv:201011929 URL <https://arxiv.org/abs/2010.11929>

- Garmong R, Bolinger R, Schumacher RS (2020) Using machine learning to improve subseasonal-to-seasonal (s2s) prediction. In: 100th American Meteorological Society Annual Meeting, AMS, URL <https://ams.confex.com/ams/2020Annual/meetingapp.cgi/Paper/369348>

- Gneiting T, Raftery AE, Westveld III AH, et al. (2005) Calibrated probabilistic forecasting using ensemble model output statistics and minimum crps estimation. *Monthly Weather Review* 133(5):1098–1118. <https://doi.org/10.1175/MWR2904.1>

- Guan H, Zhu Y, Sinsky E, et al. (2019) Systematic error analysis and calibration of 2-m temperature for the ncep gefs reforecast of the subseasonal experiment (subx) project. *Weather and Forecasting* 34(2):361–376. <https://doi.org/10.1175/WAF-D-18-0100.1>

- Hamill TM, Juras J (2006) Measuring forecast skill: Is it real skill or is it the varying climatology? *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography* 132(621C):2905–2923. <https://doi.org/10.1256/qj.06.25>

- He S, Li X, DelSole T, et al. (2020) Sub-seasonal climate forecasting via machine learning: Challenges, analysis, and advances. *arXiv preprint arXiv:2006.07972* URL <https://arxiv.org/abs/2006.07972>

- Ji L, Zhi X, Zhu S, et al. (2019) Probabilistic precipitation forecasting over east asia using bayesian model averaging. *Weather and Forecasting* 34(2):377–392. <https://doi.org/10.1175/WAF-D-18-0093.1>

- Ji L, Zhi X, Simmer C, et al. (2020) Multimodel ensemble forecasts of precipitation based on an object-based diagnostic evaluation. *Monthly Weather Review* 148(6):2591–2606. <https://doi.org/10.1175/MWR-D-19-0266.1>

- Johnstone C, Sulungu ED (2021) Application of neural network in prediction of temperature: a review. *Neural Computing and Applications* pp 1–12. <https://doi.org/10.1007/s00521-020-05582-3>
- Lorenz EN (1963) Deterministic nonperiodic flow. *Journal of atmospheric sciences* 20(2):130–141. [https://doi.org/10.1175/1520-0469\(1963\)020<0130:DNF>2.0.CO;2](https://doi.org/10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2)
- Lorenz EN (1969) Atmospheric predictability as revealed by naturally occurring analogues. *Journal of Atmospheric Sciences* 26(4):636–646. [https://doi.org/10.1175/1520-0469\(1969\)26<636:APARBN>2.0.CO;2](https://doi.org/10.1175/1520-0469(1969)26<636:APARBN>2.0.CO;2)
- Möller A, Groß J (2016) Probabilistic temperature forecasting based on an ensemble autoregressive modification. *Quarterly Journal of the Royal Meteorological Society* 142(696):1385–1394. <https://doi.org/https://doi.org/10.1002/qj.2741>
- Pegion K, Kirtman BP, Becker E, et al. (2019) The subseasonal experiment (subx): A multimodel subseasonal prediction experiment. *Bulletin of the American Meteorological Society* 100(10):2043–2060. <https://doi.org/10.1175/BAMS-D-18-0270.1>
- Peng T, Zhi X, Ji Y, et al. (2020) Prediction skill of extended range 2-m maximum air temperature probabilistic forecasts using machine learning post-processing methods. *Atmosphere* 11(8):823. <https://doi.org/10.3390/atmos11080823>
- Rasp S, Lerch S (2018) Neural networks for post-processing ensemble weather forecasts. *Monthly Weather Review* 146(11):3885–3900. <https://doi.org/10.1175/MWR-D-18-0187.1>
- Smagorinsky J (1969) Problems and promises of deterministic extended range forecasting. *Bulletin of the American Meteorological Society* 50(5):286–312. <https://doi.org/10.1175/1520-0477-50.5.286>
- Specq D, Batté L, Déqué M, et al. (2020) Multimodel forecasting of precipitation at subseasonal timescales over the southwest tropical pacific. *Earth and Space Science* 7(9):e2019EA001,003. <https://doi.org/10.1029/2019EA001003>
- Thompson PD (1957) Uncertainty of initial state as a factor in the predictability of large scale atmospheric flow patterns. *Tellus* 9(3):275–295. <https://doi.org/10.1111/j.2153-3490.1957.tb01885.x>
- Tran TTK, Bateni SM, Ki SJ, et al. (2021) A review of neural networks for air temperature forecasting. *Water* 13(9):1294. <https://doi.org/10.3390/w13091294>
- Vigaud N, Tippett MK, Yuan J, et al. (2019) Probabilistic skill of subseasonal surface temperature forecasts over north america. *Weather and Forecasting* 34(6):1789 – 1806. <https://doi.org/10.1175/WAF-D-19-0117.1>
- Vijverberg S, Schmeits M, Van der Wiel K, et al. (2020) Subseasonal statistical forecasts of eastern us hot temperature events. *Monthly Weather Review* 148(12):4799–4822. <https://doi.org/10.1175/MWR-D-19-0409.1>
- Weisheimer A, Palmer T (2014) On the reliability of seasonal climate forecasts. *Journal of the Royal Society Interface* 11(96):20131,162. <https://doi.org/10.1098/rsif.2013.1162>
- WilksDS H (2011) Statisticalmethodsintheatmosphericosciences. URL https://books.google.com.hk/books?hl=en&lr=&id=IJuCVtQ0ySIC&oi=fnd&pg=PP2&dq=Statisticalmethodsintheatmosphericosciences&ots=apJjvQAPNX&sig=09BPA0iDzM9YhSZW1pSdAwXMW9g&redir_esc=y&hl=zh-CN&sourceid=cndr#v=onepage&q&f=false
- Wulf CO, Domeisen DI (2019) Higher subseasonal predictability of extreme hot european summer temperatures as compared to average summers. *Geophysical Research Letters* 46(20):11,520–11,529. <https://doi.org/10.1029/2019GL084314>
- Xia J, Tu K, Yan Z, et al. (2016) The super-heat wave in eastern china during july–august 2013: a perspective of climate change. *International Journal of Climatology* 36(3):1291–1298. <https://doi.org/10.1002/joc.3631>

[//doi.org/10.1002/joc.4424](https://doi.org/10.1002/joc.4424)

Ye F, Hu J, Huang TQ, et al. (2021) Transformer for ei niño-southern oscillation prediction. *IEEE Geoscience and Remote Sensing Letters* <https://doi.org/10.1109/LGRS.2021.3100485>

Zhu S, Zhi X, Ge F, et al. (2021a) Subseasonal forecast of surface air temperature using superensemble approaches: Experiments over northeast asia for 2018. *Weather and Forecasting* 36(1):39–51. <https://doi.org/10.1175/WAF-D-20-0096.1>

Zhu T, Fonseca De Lima CF, De Smet I (2021b) The heat is on: how crop growth, development, and yield respond to high temperature. *Journal of Experimental Botany* 72(21):7359–7373. <https://doi.org/10.1093/jxb/erab308>

Zhu Y, Toth Z, Wobus R, et al. (2002) The economic value of ensemble-based weather forecasts. *Bulletin of the American Meteorological Society* 83(1):73–84. [https://doi.org/10.1175/1520-0477\(2002\)083<0073:TEVOEB>2.3.CO;2](https://doi.org/10.1175/1520-0477(2002)083<0073:TEVOEB>2.3.CO;2)

Zhu Y, Zhou X, Li W, et al. (2018) Toward the improvement of subseasonal prediction in the national centers for environmental prediction global ensemble forecast system. *Journal of Geophysical Research: Atmospheres* 123(13):6732–6745. <https://doi.org/10.1029/2018JD028506>