Transformer for sub-seasonal extreme high temperature probabilistic forecasting over eastern China

Wei Jin · Wei Zhang · Jie Hu · Jiazhen Chen · Bin Weng · Jianyun Gao · Tianqiang Huang

Received: 6 January 2022 / Accepted: 29 August 2022 / Published online: 9 November 2022
© The Author(s), under exclusive licence to Springer-Verlag GmbH Austria, part of Springer Nature 2022

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 experimentally. In the dataset, the multi-version data fusion method is proposed to further improve the prediction of the model with the reasonable expansion of the dataset. The performance of the 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 the 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 that happened during the third week. Overall, the deterministic forecast of T2T is superior to MLP and ECMWF due to the ability of utilizing 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 in the development of deep learning in this field and achieve the continuous progress of seamless forecasting systems.

1 Introduction

Extreme high temperature events have major impacts on human life (Zhu et al. 2021a, b). In May 2010, a persistent heat wave swept across northwest 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 research is dominated by short-range forecasting (Johnstone et al. 2021; Tran et al. 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 (Pegion et al. 2019; Vigaud et al. 2019; Vijverberg et al. 2020). As a key component of a seamless forecasting system, sub-seasonal forecasts can provide weeks of warning of extreme events.

Probabilistic forecast is essential for solving the uncertainty of the extended-range forecast. Due to a chaotic characteristic of the atmosphere and uncertainties associated with initial conditions and models (Thompson 1957; Lorenz 1963, 1969; Smagorinsky 1969), it is difficult for deterministic forecast to reproduce the future atmospheric reality. Besides, the deterministic forecast often carries a high risk and even causes irreversible losses (Boi 2004; Möller and Groß 2016). 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 forecast ability 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 (Rasp and Lerch 2018; Zhu et al. 2018; Ji et al. 2020; Zhu et al. 2021a, b). 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 (Peng et al. 2020; He et al. 2021). Among them, Peng et al. (2020) used multi-layer 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 the individual grid, thus ignoring the global correlations between grids. In the field of remote sensing, Transformer 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 Sect. 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 Sect. 5, we demonstrate the superior performance of T2T from the individual case analysis to the overall analysis of deterministic and probabilistic forecasts. Lastly, Sect. 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 6-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 by 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 6-hourly data into a daily data set, then generate 4 weeks of data in 7 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, to avoid leakage of data in the test set. The detailed division of the data set is shown in Table 1.

| Year | Dataset divided by 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 | - |

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 7 days from April to October in the training year. Then we use the 90% percentile of each grid as the extreme high temperature thresholds THR ∈ R^{H×W} (Wulff et al. 2019), as shown in Fig. 1. Our thresholds are consistent
with the high temperatures in East China from July to August 2013, indicating that the settings of the thresholds are reasonable (Xia et al. 2016).

2.3 Verification of high temperature distribution

We used the KS test to test the Gaussian distribution on a grid point-by-grid point basis for week-by-week data consisting of the same dates in the years 1995 to 2015. According to the literature (Tutorials 2018), it is known that when the P-value in the KS test is greater than 0.05, the group of data can be judged to be normally distributed. In Fig. 2, the P-values of all grid points exceeded 0.05 and, therefore, temperature follows the normal distribution.

2.4 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), 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 (Boi 2004; Hamill et al. 2006; Wilks 2011; Rasp et al. 2018; Ji et al. 2019; Specq et al. 2020).

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. 3. 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 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 in Eq. (1).

\[
\begin{align*}
\alpha_0 &= X_{\text{pre}} \cdot PD(Conv(x)) + \omega X_{\text{pos}}, \omega \in [0, 1] \\
\alpha_1 &= MHA(LN(a_{l-1})) + a_{l-1}, l = 1, \cdots, L \\
\alpha_L &= MLP(LN(a_L)) + a_L, l = 1, \cdots, L \\
y &= RS(RS(a_L)) 
\end{align*}
\] (1)

Firstly, our input is the previously generated \( x \in \mathbb{R}^{H_0 \times W_0} \). After one convolution operation (Convolution, Conv), the feature map \( f_1 \in \mathbb{R}^{C_1 \times H_1 \times W_1} \) is obtained. The feature map is then divided (Patch Division, PD) into \( N \) smaller feature maps of the same size \( f_2 \in \mathbb{R}^{C_1 \times P \times P} \) and flattened to one-dimensional vectors. Finally, the semantic embedding \( f_3 \in \mathbb{R}^{N \times C_2} \) 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 \( X_{\text{pre}} \in \mathbb{R}^{1 \times C_2} \). We have also used position embedding denoted by \( X_{\text{pos}} \in \mathbb{R}^{N+1 \times C_2} \), 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 \( \omega \) 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 multi-layer perceptron (MLP) and the layer normalization (LN).
Among them, multi-head attention is composed of multiple self-attention structures, which can extract the whole image feature and has better global feature extraction ability. After multiple encodings, our forecast result is output in the form of a tensor \( y \in \mathbb{R}^{1 \times H_0 \times W_0} \) 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 et al. 2018). To establish the association from deterministic to probabilistic forecasting, we enter \( y \) as the expectation \( \mu \) of a Gaussian distribution into the probabilistic forecasting model \( T2T_P \), and generate the standard deviation \( \sigma \) under a point-by-point 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 \( \sigma \) and the expectation \( \mu \), 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 Eq. (2).

\[
y = T2T_C(x) \\
\mu = y \\
\sigma = T2T_P(\mu) \\
F(THR) = P(X > THR) = \frac{1}{\sqrt{2\pi}\sigma} \int_{THR}^{\infty} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt
\]

### 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 Eqs. (3) and (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), the number of heads of attention mechanism is set to 2 and 8, respectively. The main experimental section has two models whose results are compared with the ensemble mean of the ECMWF: multi-layer perceptron (MLP) and Transformer to High Temperature (T2T), where MLP directly uses the network structure of the reference (Peng et al. 2020).

\[
Loss = \frac{1}{MN} \sum_{i=1}^{M} \sum_{g=1}^{N} \left| X_{i,g} - Y_{i,g} \right|
\]
In deterministic forecasting, \( x \) is the forecast value and \( Y \) is the observed value (Ji et al. 2019).

\[
\begin{align*}
    m &= \frac{Y - \mu}{\sigma}, \\
    \text{CRPS}(t) &= \sigma \left( m(2\Phi(m) − 1) + 2\varphi(m) − \frac{1}{\sqrt{2\pi}} \right) \quad (4)
\end{align*}
\]

\[
\text{Loss} = \frac{1}{MN} \sum_{i=1}^{M} \sum_{k=1}^{N} \text{CRPS}(N_{i,k}(\mu, \sigma^2), Y_{i,k})
\]

In probabilistic forecasting, \( Y \) is the observed value; \( \Phi \) and \( \varphi \) are the predictive cumulative distribution function (CDF) and the probability density function (PDF) of a normal distribution with \( N(\mu, \sigma^2) \) (Hamill et al. 2006).

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. To verify whether the existence of Transformer plays a positive role in the model, we remove the Transformer in T2T to form a new model, denoted 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 a positive role in the model, which makes the forecast results more meaningful for an extended period.

In Eq. (1), we add a weight \( \omega \) to the use of position embedding, where we set the weights \( \omega \) 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 indicate that the existence of position embedding in our task does not effectively improve the forecast performance of the model, but rather has a greater negative impact on the forecast results.

In summary, when \( \omega \) 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 sacrifice the accuracy of the data by using position

**Table 2** The influence of Transformer

| Indicators | Models    | Week 1 | Week 2 | Week 3 | Week 4 |
|------------|-----------|--------|--------|--------|--------|
| 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   |

T2T (T2T-NT) means the model with (without) Transformer.

**Table 3** Results of indicators with different weights \( \omega \)

| \( \omega \) | Week 1 | Week 2 | Week 3 | Week 4 |
|-------------|--------|--------|--------|--------|
| MAE         | 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        | 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         | 0.55   | 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         | 0.96   | 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          | 0.04   | 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   |

The bolded part is the best result in each group of comparison experiments. MAE, RMSE, ETS, and ACC are deterministic forecast indicators; BS is probabilistic forecast indicator.
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. 4, 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, the 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 4-week forecast initiated from 2 June 2018 and exhibit the spatial distribution of errors, with the
domain of China experiencing a 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. 5. Compared with the absolute error of ECMWF, the errors of MLP are significantly reduced and lighter in many places, which indicate that the multi-layer perceptron works well in the revision problem of multi-grid regression and has largely been able to effectively revise the ECMWF forecast data. Furthermore, from the first week to the fourth week, the percentage of better grid points of T2T compared to MLP gradually increases, while the overall forecast error is smaller, which indicates that the effect of T2T improves significantly with a longer extension period. It is worth noting that the week 3 forecast of ECMWF is slightly worse than week 4 found in previous studies (Wulff et al. 2019). The forecasting method in this paper is based on the ECMWF forecasting results, and it is consistent with the ECMWF in terms of the overall trend and fluctuation pattern of the forecasting effect, which precisely shows that our model learns some features of the ECMWF forecasts.

In the analysis of the probabilistic forecasts, we compare the probabilistic forecasts of the three methods with the observations, as shown in Fig. 6. The first row in red marks the real high temperature areas for 4 consecutive weeks. In the ECMWF forecasts, the first 2 weeks are good, with high probability values above 95% for some of the high temperature areas, but the second 2 weeks are poor, with little correlation with the true high temperature areas even at the maximum probability values. In the MLP forecast is added in the lower right corner of each subgraph, and the percentage of grid points that T2T is superior to MLP is calculated as 46.1%, 57.9%, 65.7%, and 60.6%, respectively.
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 2 weeks and a weak correlation with observations at the 50% highest probability in the second 2 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 perform well. Larger high temperature fallout areas and high probability values above 95% are forecast in the first 2 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 2 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 real 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. 7, we compare the results of the three methods. The MLP outperforms the ECMWF in all metrics, indicating that the multi-layer 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, we show the spatial distribution of TCC for the three methods in Fig. 8. Comparing the ECMWF with the MLP, we find that the former is slightly more effective than the latter in the latter 2 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 in 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.

![Histogram of deterministic forecast evaluation. MAE, RMSE, ETS, ACC are the four indicators](image)
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. Ultimately, the probabilities are calculated from the distributions. Since our study focus on the dichotomous extreme event and CRPS measures the overall probabilistic performance, it is not employed here.

Figure 9 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 performs 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. 10. 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 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. 10, we find that a grid point located near the Taihang Mountains consistently exhibits the 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 et al. 2014). In addition to the skill of forecast, we analyze the reliability of forecast as in Fig. 11. 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, and 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. 11 represent different forecasting reliability. Lines above the diagonal line exhibit under-forecasting bias while below the 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. 11) 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 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.
Author contribution WJ and WZ did the experiments, writing, and analysis together, and the other authors helped with touch-ups and other content refinements.

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

Data availability 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.

Declarations

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent to publication Not applicable.

Conflict of interest The authors declare no competing interests.

References

Boi P (2004) Probabilistic temperature forecast by using ground station measurements and ECMWF ensemble prediction system. Meteorol Appl 11(4):301–309. https://doi.org/10.1017/S1350482704001380
Dosovitskiy A et al (2020) An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929. https://arxiv.org/abs/2010.11929. Accessed 14 Nov 2021
Garmong R et al (2020) Using machine learning to improve subseason-to-seasonal (S2S) prediction. 100th American Meteorological Society Annual Meeting, AMS. https://ams.confex.com/ams/2020Annual/meetingapp.cgi/Paper/369348. Accessed 14 Nov 2021
Gneiting T et al (2005) Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. Mon Weather Rev 133(5):1098–1118. https://doi.org/10.1175/MWR2904.1
Guan H et al (2019) Systematic error analysis and calibration of 2-m temperature for the NCEP GEFS reforecast of the Subseasonal Experiment (SubX) Project. Weather Forecast 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? Q J R Meteorol Soc: A J Atmos Sci, Appl Meteorol Phys Oceanogr 132(621C):2905–2923. https://doi.org/10.1256/qj.06.25
He S et al (2021) Sub-seasonal climate forecasting via machine learning: challenges, analysis, and advances. Proceedings of the AAAI Conference on Artificial Intelligence. https://arxiv.org/abs/2006.07972
Ji L et al (2019) Probabilistic precipitation forecasting over East Asia using Bayesian model averaging. Weather Forecast 34(3):377–392. https://doi.org/10.1175/MWR-D-19-0266.1
Ji L et al (2020) Multimodel ensemble forecasts of precipitation based on an object-based diagnostic evaluation. Mon Weather Rev 148(6):2591–2606. https://doi.org/10.1175/WAF-D-18-0093.1
Johnstone C, Sulungu ED (2021) Application of neural network in prediction of temperature: a review. Neural Comput Appl 33(18):11487–11498. https://doi.org/10.1007/s00521-020-05582-3
Lorenz EN (1963) Deterministic nonperiodic flow. J Atmos Sci 20(2):130–141. https://doi.org/10.1175/1520-0469(1963)020<0130:DNP%3E2.0.CO;2
Lorenz EN (1969) Atmospheric predictability as revealed by naturally occurring analogues. J Atmos Sci 26(4):636–646. https://doi.org/10.1175/1520-0469(1969)026<0636:APARM%3E2.0.CO;2
Müller A, Groll J (2016) Probabilistic temperature forecasting based on an ensemble autoregressive modification. Q J R Meteorol Soc 142(696):1385–1394
Pegion K et al (2019) The Subseasonal Experiment (SubX): a multimodel subseasonal prediction experiment. Bull Am Meteor Soc 100(10):2043–2060. https://doi.org/10.1175/BAMS-D-18-0270.1
Peng T 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 postprocessing ensemble weather forecasts. Mon Weather Rev 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. Bull Am Meteor Soc 50(5):286–312. https://doi.org/10.1175/1520-0477-50.5.286
Speq D et al (2020) Multimodel forecasting of precipitation at subseasonal timescales over the southwestern tropical Pacific. Earth and Space Science 7(9) e2019EA001003. 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 et al (2021) A review of neural networks for air temperature forecasting. Water 13(9):1294. https://doi.org/10.3390/w13091294
Tutorials S (2018) SPSS Kolmogorov–Smirnov test for normality. https://www.spss-tutiorials.com/spss-kolomgorov-smirnov-test-for-normality/. Accessed 14 Nov 2021
Vigaud N et al (2019) Probabilistic skill of subseasonal surface temperature forecasts over North America. Weather Forecast 34(6):1789–1806. https://doi.org/10.1175/WAF-D-19-0117.1
Vijverberg S et al (2020) Subseasonal statistical forecasts of eastern US hot temperature events. Mon Weather Rev 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. J R Soc Interface 11(96):20131162. https://doi.org/10.1098/rsif.2013.1162
Wilks DS (2011) Statistical methods in the atmospheric sciences. Academic Press
Wulff CO, Domeisen DI (2019) Higher subseasonal predictability of extreme hot European summer temperatures as compared to average summers. Geophys Res Lett 46(20):11520–11529. https://doi.org/10.1029/2019GL084314

Xia J et al (2016) The super-heat wave in eastern China during July–August 2013: a perspective of climate change. Int J Climatol 36(3):1291–1298. https://doi.org/10.1002/joc.4424

Ye F et al (2021) Transformer for El Niño-Southern oscillation prediction. IEEE Geosci Remote Sens Lett 19:1–5. https://doi.org/10.1109/LGRS.2021.3100485

Zhu Y et al (2002) The economic value of ensemble-based weather forecasts. Bull Am Meteor Soc 83(1):73–84. https://doi.org/10.1175/1520-0477(2002)083<0073:TEVOEB>2.0.CO;2

Zhu Y 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

Zhu S et al (2021) Subseasonal forecast of surface air temperature using superensemble approaches: Experiments over Northeast Asia for 2018. Weather Forecast 36(1):39–51. https://doi.org/10.1175/WAF-D-20-0096.1

Zhu T et al (2021) The heat is on: how crop growth, development, and yield respond to high temperature. J Exp Bot 72(21):7359–7373. https://doi.org/10.1093/jxb/erab308

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.