Atmospheric Research 330 (2026) 108602

Contents lists available at ScienceDirect

Atmospheric Research

journal homepage: www.elsevier.com/locate/atmosres

Sub-seasonal prediction of extreme high temperature over the Yangtze River Basin based on deep learning

Shifeng Pan [a,b,*], Zhicong Yin [a,b], Yi Fan [a,b], Tingting Han [a,b], Mingkeng Duan [a,b], Huijun Wang [a,b,c]

[a] State Key Laboratory of Climate System Prediction and Risk Management/Key Laboratory of Meteorological Disaster, Ministry of Education/Collaborative Innovation Center on Forecast and Evaluation of Meteorological Disasters, Nanjing University of Information Science and Technology, Nanjing 210044, China
[b] School of Atmospheric Sciences, Nanjing University of Information Science and Technology, Nanjing 210044, China
[c] Nansen-Zhu International Research Centre, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing 100029, China

ARTICLE INFO

Keywords:
Sub-seasonal
Res-UNet
Extreme high temperature
Deterministic forecast
Probabilistic forecast

ABSTRACT

Extreme high temperatures frequently affect the densely populated Yangtze River Basin (YRB) in China, impacting livelihoods and ecosystems. However, predicting such events on a sub-seasonal scale remains challenging. This study develops an advanced deterministic and probabilistic forecasting framework based on the residual U-Net (Res-UNet) architecture, trained and validated from 2003 to 2018 and tested over 2019–2022 over the YRB. In deterministic forecast, Res-UNet substantially improves temporal correlation coefficient by 14 % compared to the European Center for Medium-range Weather Forecasts (EC) and empirical quantile mapping (QM) at the 3–4 weeks lead time. Notably, it achieves a probability of detection (POD) of 0.3, maintaining predictive skill where other models falter. For the probabilistic forecast, Res-UNet outperforms EC and QM at longer lead times, providing Brier Skill Score (BSS) and Continuous Ranked Probability Skill Score (CRPSS), with lower Brier Scores (BS), especially in Chongqing and Sichuan. During the extreme events of 2022, Res-UNet delivers high-confidence predictions of extreme high temperatures up to 3 weeks in advance. Furthermore, outperforming other machine learning models highlights its potential in enhancing sub-seasonal extreme high-temperature forecast.

1. Introduction

Extreme high temperatures have profound impacts on human life (Zhu et al., 2021). During the summer of 2022, Europe and Asia faced unprecedented heatwaves. In the United Kingdom, temperatures surpassed 40 °C for the first time ever, leading to at least 15,000 heat-related fatalities in Europe (Yin et al., 2023). In Asia, China experienced the most widespread and prolonged heatwave on record during the summer of 2022 (from mid-June to late August), marked by exceptionally intense summer heatwaves and drought in the Yangtze River Basin (YRB) (Lu et al., 2023; Sun et al., 2023; Yin et al., 2024). Specifically, in Beipei, Chongqing, temperature soared to 45 °C (Hua et al., 2023). Moreover, industries such as public health and water resource management require reliable and actionable predictions on sub-seasonal scale (Vitart, 2017). This underscores the urgent need for more reliable methods to predict extreme high temperatures in advance, aiming to mitigate both economic and human losses.

Current weather forecast depends on numerical weather prediction models (Lorenc, 1986). However, uncertainties in the initial conditions of these models, along with the inherently chaotic nature of the atmosphere, introduce biases into numerical forecast results (Edward, 1963), which in turn limits the accuracy of deterministic forecasts. Based on this, probabilistic forecast has emerged, shifting forecast results from a single deterministic value to a probabilistic outcome. Probabilistic forecast seeks to capture the uncertainty associated with a variable by providing complete forecasting probability distributions instead of merely a single value (Gneiting and Katzfuss, 2014). Therefore, probabilistic forecast provides an effective way to quantify the uncertainty of extreme predictions.

Numerous post-processing approaches have been developed to enhance probabilistic forecast, complementing the progress of numerical weather prediction models. Since the 1990s, agencies in the US and Japan have applied Model Output Statistics (MOS) for operational forecast (Toth, 2001). In addition, Bayesian Model Averaging (BMA) has

* Corresponding author.
E-mail address: psf@nuist.edu.cn (S. Pan).

https://doi.org/10.1016/j.atmosres.2025.108602
Received 6 March 2025; Received in revised form 20 October 2025; Accepted 29 October 2025
Available online 1 November 2025

been widely adopted to extend probabilistic forecast from temperature to variables like precipitation, improving accuracy (Raftery et al., 2005; Sloughter et al., 2007; Ma et al., 2016). For example, Kolachian and Saghafian (2019) showed BMA enhanced precipitation forecasts using European Centre for Medium-Range Weather Forecasts (EC). However, probabilistic forecast faces calibration issues, with forecast probabilities often not matching observed frequencies—for example, events predicted with an 80 % probability tend to occur less frequently than expected (Hamill and Colucci, 1997; Wilks, 2013). Moreover, both over-confidence and under-confidence are common, especially in forecasts of extreme events, which reduces forecast reliability (Wilks, 2013). Additionally, most studies focus on short to medium-term forecasts, with relatively little attention given to probabilistic forecast at the sub-seasonal timescale (Vigaud et al., 2018).

The timescale of sub-seasonal forecast, typically covering 10–30 days (Sun et al., 2010), falls between weather and seasonal forecast. At this timescale, atmospheric initial condition signals have largely dissipated and external forcing signals have yet to fully emerge, often making it a "predictability desert" due to its reliance on fluctuating climatic factors (Vitart et al., 2017). Despite advances in sub-seasonal to seasonal (S2S) modeling (Pegion et al., 2019), forecast skill tends to deteriorate with longer lead times (Zhu et al., 2023). Previous studies have shown that the forecast skill of S2S models for extreme high temperature over the YRB in China declines significantly beyond a 2-week lead time (Xie et al., 2020; Jin et al., 2022; Lei et al., 2022; Tang et al., 2023). Pyrina and Domeisen (2023) demonstrated that EC predicted heatwave onset and duration well at 1-week lead, but skill dropped at 2–3 weeks. Notably, Lin et al. (2022) found that while most S2S models predicted the 2021 western North America heatwave 2–3 weeks ahead, they significantly underestimated its intensity (Vitart et al., 2025). Overall, current S2S models still face challenges in the sub-seasonal prediction of extreme high temperature.

Thus, to improve the accuracy of regional S2S forecasts, a variety of traditional methods have been proposed and applied, including the interannual increment approach (Fan et al., 2008), quantile mapping (QM) (Tong et al., 2021), spatiotemporal projection models (Hsu et al., 2020), dynamical forecasting methods (Guo et al., 2017), and hybrid dynamical-statistical approaches (White et al., 2017). These methods primarily aim to correct systematic biases in raw model outputs, thereby improving forecast accuracy in climate prediction. For example, QM typically achieves rapid bias correction by fitting the historical probability distribution, offering the advantage of quick adjustment. However, QM generally operates at individual grid points, lacking consideration of spatial coherence and structural dependencies (Tong et al., 2021). In contrast, dynamical–statistical models combine the strengths of physical mechanisms and statistical methods, which improves predictive skill to some extent (Hsu et al., 2020). However, these models still struggle to capture nonlinear features.

With the advancement of machine learning, particularly deep learning, significant progress has been made in the field of S2S prediction, largely due to its effectiveness in capturing complex nonlinear relationships in data (Rasp and Lerch, 2018; Ding et al., 2024). Zhang et al. (2023) employed a random forest (RF) with multivariable post-processing to enhance S2S prediction of extreme precipitation over the continental United States. Similarly, Zhou and Liu (2025) utilized a convolutional neural network (CNN) trained on reanalysis data to link East Asian summer monsoon precipitation with circulation fields, achieving improved correction at a 1-week lead time. Weyn et al. (2021) developed a large-scale deep learning ensemble weather prediction system capable of producing 4–6 weeks temperature forecast comparable to the EC. Furthermore, Xie et al. (2024) combined CNN with multiscale precursor signals to enhance surface air temperature over China for lead times of 10–30 days. However, these machine learning approaches struggle to capture the complex spatial structures, and their skills typically declined sharply beyond a 2-week lead time, particularly for extreme events (Jin et al., 2022; He et al., 2021). These limitations underscore the urgent need for model architectures that can effectively capture multiscale spatial patterns and sustain strong accuracy at extended lead times.

Among deep learning architectures, the U-Net model stands out for its ability to capture multi-dimensional geographic features and reduce errors, leading to significant improvements in sub-seasonal forecast (Vitart et al., 2022; Zhu et al., 2022; Lyu et al., 2023). Specifically, a U-Net based architecture integrating numerical model outputs with multiple atmospheric variables has markedly enhanced the prediction of summer extreme precipitation in southern China (Lyu et al., 2023), suggesting its strong potential for subseasonal applications. Hence, this study introduces a U-Net architecture to enhance the prediction skill of extreme high temperature at lead times of 3 and 4 weeks.

In this study, the YRB is a hotspot for extreme high-temperature events, where frequent and intense heatwaves severely affect densely populated and economically developed areas (Hsu et al., 2017). Given that sub-seasonal forecast mainly targets prediction beyond two weeks, this study specifically focuses on week 3 and 4 predictions. However, at these timescales, deterministic forecast exhibits substantial errors (Tong et al., 2021), and probabilistic forecast shows limited effectiveness for extreme high temperature over the YRB (Wilks, 2013). Therefore, we propose a dual deep learning model based on the U-Net architecture for deterministic and probabilistic forecast, and evaluates its performance.

2. Data and methods

2.1. Data

The EC model through S2S Project version 2023 (Vitart et al., 2017) is launched twice a week, featuring a horizontal resolution of $1° \times 1°$. EC provides reforecast datasets for the maximum temperature (tasmax) at 2 m along with other factors such as geopotential height (Z), wind components (U, V), vertical velocity (W), specific humidity (Q), and temperature (T) at the 200, 500, and 850 hPa pressure levels (Table 1). The selected atmospheric variables are linked to the generation of high temperature, including geopotential height and winds that trigger heatwaves through subsidence and adiabatic heating (Lu et al., 2023). Additionally, specific humidity, which aligns with temperature trends, plays a key role in capturing climate signals relevant to sub-seasonal forecast (De Bruin et al., 1999). We employ reforecast datasets with prediction lead times spanning 0 to 28 days for the months of June to August from 2003 to 2022. And the other data consists of daily tasmax provided by the Fifth Generation European Center for Medium-Range Weather Forecasts Reanalysis Dataset (ERA5) (Hersbach et al., 2020).

The study period spans from June to August, covering the years 2003 to 2022. The data is categorized into three segments: a training set (2003–2015), a validation set (2016–2018), and a testing set (2019–2022). Our training area includes part of China with a spatial range of 21°–36°N and 103°–122°E, mainly focusing on the Yangtze River Basin (YRB; 25°N–34°N, 103°E–120°E). A daily maximum temperature of 35 °C is widely recognized for its significant impacts on health and ecosystems (Tan et al., 2007). Studies indicate that the 95th percentile of summer maximum temperatures in the YRB falls between 34 and 36 °C (Li and Huang, 2011), supporting the use of 35 °C as a

Table 1
The input variables for the Res-UNet model.

	Variable name	Time	Level
input	maximum temperature		2 m
	geopotential height		200/500/850 hPa
	zonal wind	Total: 2003–2022	200/500/850 hPa
	meridional wind	Training: 2003–2015	200/500/850 hPa
	vertical velocity	Validation: 2016–2018	200/500/850 hPa
	temperature	Testing: 2019–2022	200/500/850 hPa
	specific humidity		200/500/850 hPa
label	maximum temperature		2 m

representative threshold. This also aligns with the China Meteorological Administration's official high-temperature warning standard, defining extreme heat as days exceeding 35 °C (Tan et al., 2007; Li and Sun, 2018). Therefore, this study adopts 35 °C as the threshold for identifying extreme high temperature.

2.2. Res-UNet model

In this study, we utilize a U-Net framework with the flowchart shown in Fig. 1. The model incorporates residual layers, downsampling layers, deconvolution layers, and shortcut connections. Unlike the original U-Net by Ronneberger et al. (2015), our model replaces standard convolutional blocks with residual blocks (Res-UNet) to reduce overfitting and enhance learning capacity. Using this framework, we develop both deterministic and probabilistic forecast model.

Specifically, the architecture of Res-UNet (Fig. 1) processes input features with dimensions (16, 19, 20), representing the width and height of the study area and the number of input channels (Fig. 1, blue shading). These channels include normalized 2-m tasmax forecasts and atmospheric variables from EC. Two residual blocks ensure consistent input and output dimensions, followed by two max-pooling layers to reduce spatial dimensions. Up-sampling layers restore the spatial resolution, while skip connections preserve down-sampling information. The final deterministic forecast matches the spatial dimensions of the input temperature forecast (Pan et al., 2024).

Subsequently, the forecasts from individual EC ensemble members are used as inputs to the probabilistic model, which estimates the standard deviation (σ) of the predictive distribution (Fig. 1, orange shading). Following previous studies (Gneiting et al., 2005; Pan et al., 2022), high temperatures are assumed to approximately follow a Gaussian distribution, with the deterministic forecast serving as the expected value (μ) of the distribution. The probabilistic forecast is then generated using both the expected value and standard deviation. Finally, the probability of exceeding a predefined extreme threshold (35 °C) is calculated by integrating the Gaussian probability density function beyond that threshold. The same network architecture is used for both deterministic and probabilistic forecast.

To enhance the performance of deep learning models in forecasting extreme events, incorporating a weighted loss function has proven to be a critical strategy (Hu et al., 2021). Given our attention to extreme high-temperature events, we design the loss function based on the extreme threshold of 35 °C. Consequently, we construct a weighted function that incorporates the extreme threshold, mean square error, threat score for effectively capturing deterministic forecast (Text S1; Pan et al., 2024). For probabilistic forecast, Continuous Ranked Probability Score (CRPS) is selected as the loss function to measure the disparity between predicted and observed distributions. The CRPS is defined as follows:

$$CRPS(P,F) = \int_{-\infty}^{\infty} (F(x) - 1\{y \le x\})^2 dx \quad (1)$$

Here, P refers to cumulative distribution function (CDF) of observation, F denotes predicted CDF in model, and $1\{y \le x\}$ serves as an indicator function (equivalent to the CDF of a deterministic value) that equals 1 when $\{y \le x\}$ and 0 otherwise.

Based on the Res-UNet model described above, the output of the Res-UNet is the daily high temperature within a timeframe of 1 to 4 weeks, using June 1 as the target date (Fig. S1). The forecasts are initialized on May 25, May 18, May 11, and May 4, which correspond to lead times of 1–4 weeks, respectively. Each initial date forecasts data for the next 46 days, focusing on summer outputs for evaluation. As forecast lead times

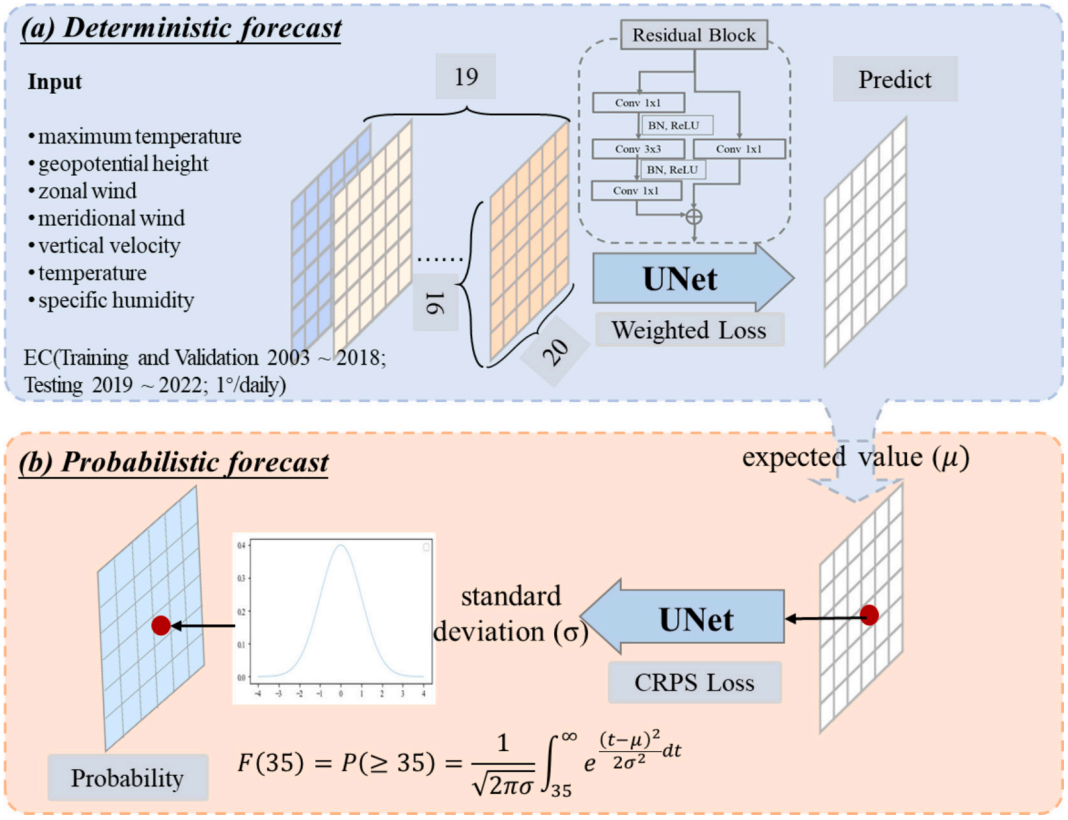

Fig. 1. Architecture of a novel hybrid forecasting approach integrates deterministic and probabilistic predictions using the residual U-Net (Res-UNet). (a) Flowchart of deterministic forecast (blue shading), including input variables and other information. (b) Flowchart of the probabilistic forecast (orange shading), where F and P denote the predicted cumulative distribution function (CDF) and probability density function (PDF), respectively. Extreme high temperatures are defined using a threshold of 35 °C, and the expected value (μ) and standard deviation (σ) are also illustrated. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

increase, the forecast error tends to grow (Liang and Lin, 2018). Moreover, sub-seasonal forecast primarily targets prediction skills for periods longer than 2 weeks. Therefore, we only train separate Res-UNet models for fixed lead times of 3 weeks and 4 weeks for predictions. The model's performance is particularly assessed for predicting extreme high temperatures within 3–4 weeks lead times. Hyperparameters of Res-UNet are selected based on validation performance, aiming to balance overall and extreme prediction. Training employs the Adam optimizer with a dynamic learning rate schedule. Key parameters, including learning rate (10^{-4}), batch size (8), and 100 epochs, are tuned to optimize performance at a fixed lead time during training.

2.3. Metrics

For deterministic forecast, performance is assessed using temporal correlation coefficient (TCC) and root mean square error (RMSE) (Text S2). To evaluate the deterministic forecast on extremes, the categorical verification scores are calculated (Table S1), such as the probability of detection (POD), equitable threat score (ETS), heidke skill score (HSS), and false alarm ratio (FAR), as follows:

$$ETS = \frac{\left(a - \frac{(a+b)(a+c)}{a+b+c+d}\right)}{\left(a + b + c - \frac{(a+b)(a+c)}{a+b+c+d}\right)} \tag{2}$$

$$POD = \frac{a}{a+c} \tag{3}$$

$$HSS = \frac{2(a*d - b*c)}{(a+c)(c+d) + (a+b)(b+d)} \tag{4}$$

$$FAR = \frac{b}{a+b} \tag{5}$$

in which a, b, c and d indicate the samples of hit, false alarm, miss and correct rejection, respectively. The range of ETS is from $-\frac{1}{3}$ to 1, and POD is 0 to 1, while the HSS score ranges from $-\infty$ to 1. Lower FAR values indicate higher forecasting skills.

For the evaluation of probabilistic prediction, we utilize the area under the curve (AUC) of receiver operating characteristics (ROC) to assess the capability for identifying extreme and non-extreme events. This curve plots the hit rate against the false alarm ratio for predictions. Additionally, we utilize Brier Score (BS) and Brier Skill Score (BSS) to evaluate the probabilistic forecast performance of the model. The BS quantifies the forecast error, The BS is calculated as follows:

$$BS = \frac{1}{N}\int_{n=1}^{N} (f_n - O_n)^2 \tag{6}$$

Here, N represents the number of forecast events, O_n indicates that the extreme high-temperature event occurs at time n, equaling 1; otherwise, it equals 0. f_n represents the forecast probability of the event. The BS is a negative-oriented metric, ranging from 0 to 1.

The BSS measures the improvement of the forecast compared to a baseline climatological forecast. The BSS is:

$$BSS = 1 - \frac{BS_f}{BS_{ref}} \tag{7}$$

Where BS_{ref} refers to the climatological probabilistic forecast, while BS_f represents predictive probability. If the *BSS* is 0, it means that the score is equivalent to the climate forecast; if the *BSS* is greater than 0, it indicates that the forecast has some predictive value; conversely, if the *BSS* is less than 0, it means the forecast is less meaningful than the climate probability and has no predictive value.

The Area Under the Curve (AUC) refers to the area under the ROC curve and is used to evaluate the performance of a classification model. The AUC value ranges from 0.5 to 1, with higher values indicating better model performance. The AUC can be computed through the integral of False Positive Rate (TPR) over False Positive Rate (FPR) on the ROC curve (Text S2).

$$AUC = \int_0^1 TPR\, d(FPR) \tag{8}$$

Continuous Ranked Probability Skill Score (CRPSS)

$$CRPSS = 1 - \frac{CRPS_f}{CRPS_{ref}} \tag{9}$$

Where $CRPS_{ref}$ denotes the climatological probabilistic forecast, and $CRPS_f$ represents the forecasted probability distribution. A $CRPSS = \mathbf{1}$ indicates a perfectly accurate probabilistic forecast; $CRPSS = \mathbf{0}$ means the forecast performs equivalently to the reference forecast; and $CRPSS < \mathbf{0}$ implies that the forecast performs worse than the reference forecast. Detailed information on evaluation metrics can be found in Table 2 and Text S2.

2.4. Shapley additive explanations

Shapley Additive Explanations (SHAP), introduced by Lundberg and Lee (2017), is a method designed to interpret the outputs of machine learning models, providing a more reliable way to explain black-box models. The SHAP method leverages the classic Shapley value from game theory to connect optimal credit allocation with local explanations. By decomposing a model's output into the sum of individual feature impacts, SHAP enables a better understanding of the importance of each feature, ultimately supporting more informed decision-making.

2.5. Evaluation strategy for sub-seasonal extreme high temperature

In this study, we first evaluate the deterministic forecasting

Table 2
Deterministic and probabilistic forecasting skill scores.

Category	Evaluation metrics	Calculation	Skillful Range
Deterministic	RMSE (Root Mean Square Error)	Eq. (4) in Text S2	[0, +∞), the smaller the better
	TCC (Temporal Correlation Coefficient)	Eq. (5) in Text S2	(0,1), the larger the better
	ETS (Equitable Threat Score)	Eq. (2)	(−1/3, 1), the larger the better
	POD (Probability of Detection)	Eq. (3)	(0, 1), the larger the better
	HSS (Heidke Skill Score)	Eq. (4)	(−∞, 1), the larger the better
	FAR (False Alarm Ratio)	Eq. (5)	[0, +∞), the smaller the better
	BS (Brier Score)	Eq. (6)	[0, +∞), the smaller the better
	BSS (Brier Skill Score)	Eq. (7)	(0, 1), the larger the better
Probabilistic	ROC (Receiver Operating Characteristics)	Eqs. (6) and (7) in Text S2	–
	AUC (Area under the ROC curve)	Eq. (8)	(0.5, 1), the larger the better
	CRPSS (Continuous Ranked Probability Skill Score)	Eq. (9)	(−∞, 1), the larger the better

performance of Res-UNet model for summer extreme high temperatures during 2019–2022, in comparison with EC and QM at 3–4 weeks lead times. Deterministic forecast metrics—RMSE, TCC, ETS, POD, HSS, and FAR (introduced in Section 2.3)—are used to quantitatively assess the accuracy and skill. In addition, the QM method is used for comparison with Res-UNet. During training, CDFs of observed and predicted high temperatures are matched, and this mapping is applied to the testing set (2019–2022).

At the sub-seasonal timescale, deterministic forecast offers limited reference value, particularly for extreme events. Probabilistic forecast can quantify uncertainty and provide more reliable information to support risk assessment and decision-making. In this study, we evaluate the probabilistic forecasts from Res-UNet, compared to EC and QM using the 2019–2022 testing set. In particular, the extreme high-temperature events in 2022 over the YRB, which greatly affect people's lives (Lu et al., 2023), are too severe to be accurately predicted, leading to substantial deterministic errors of that year. Accordingly, we pay particular attention to analyzing the probabilistic forecast in 2022, presenting the spatial distribution of skill scores from EC, QM, and Res-UNet. Several probabilistic metrics are introduced, including BSS, CRPSS, ROC, AUC, BS, and reliability analysis, providing a thorough assessment of the Res-UNet's probabilistic forecasting performance.

For additional context, Res-UNet is also compared with several other machine learning methods in Discussion, including artificial neural network (ANN), convolutional long-short term memory (ConvLSTM), CNN, and RF. These models are chosen to represent different modeling strategies. Specifically, ANN captures nonlinear relationships without spatial features, CNN extracts spatial patterns for comparison with the residual U-Net, ConvLSTM captures spatiotemporal dependencies, and RF serves as a traditional ensemble method. All models are trained using optimized algorithms and hyperparameters (see Table S2).

3. Results

3.1. Deterministic forecast

Deterministic forecast aims to predict specific high temperature values, characterized by their uniqueness and absolute nature. In this study, the variable data (Table 1) from 2003 to 2015 is input into the Res-UNet framework for training, followed by validation using data from 2016 to 2018. And finally, the deterministic forecast results for high temperatures are evaluated on the testing set from 2019 to 2022. Therefore, we evaluate the summer high temperature output from EC, QM, and Res-UNet for 3–4 weeks lead times in 2019–2022 averaged over the YRB (Fig. 2). Within this period, Res-UNet's skill surpasses that of EC and QM, with improvements in both TCC and reductions in RMSE (Fig. 2a, b). This indicates that Res-UNet has higher predictive accuracy and trend-capturing ability over longer forecast periods (Fig. 2a, b, S2). However, the performance of EC and QM in terms of TCC is essentially the same, indicating that QM has limited calibration capabilities in this regard, which aligns with findings from previous studies (Li et al., 2022). Remarkably, at lead time of 3–4 weeks, TCC increases by approximately 14 % while RMSE decreases by 2 % in Res-UNet.

For the deterministic forecast of the summer of 2022, the results demonstrate that Res-UNet better captures the spatial distribution of observations (Fig. S2, S3). At the 3–4 weeks lead times, Res-UNet outperforms EC and QM in capturing nonlinear processes. Res-UNet can accurately predict the high-temperature areas in observation, with slightly higher intensity. Compared to EC and QM, the Res-UNet shows higher spatial correlation with the actual observation (Fig. S3). Moreover, extreme high temperatures in the Chongqing and Sichuan regions are predicted, whereas EC and QM fail to predict these. Additionally, Res-UNet exhibits higher TCC and lower errors relative to the QM and EC, suggesting a significant improvement over the Lower Yangtze River (Fig. S4, S5). These results preliminarily highlight the superiority of the Res-UNet to EC and QM in deterministic forecast.

Furthermore, to assess EC, QM, and Res-UNet in predicting extreme

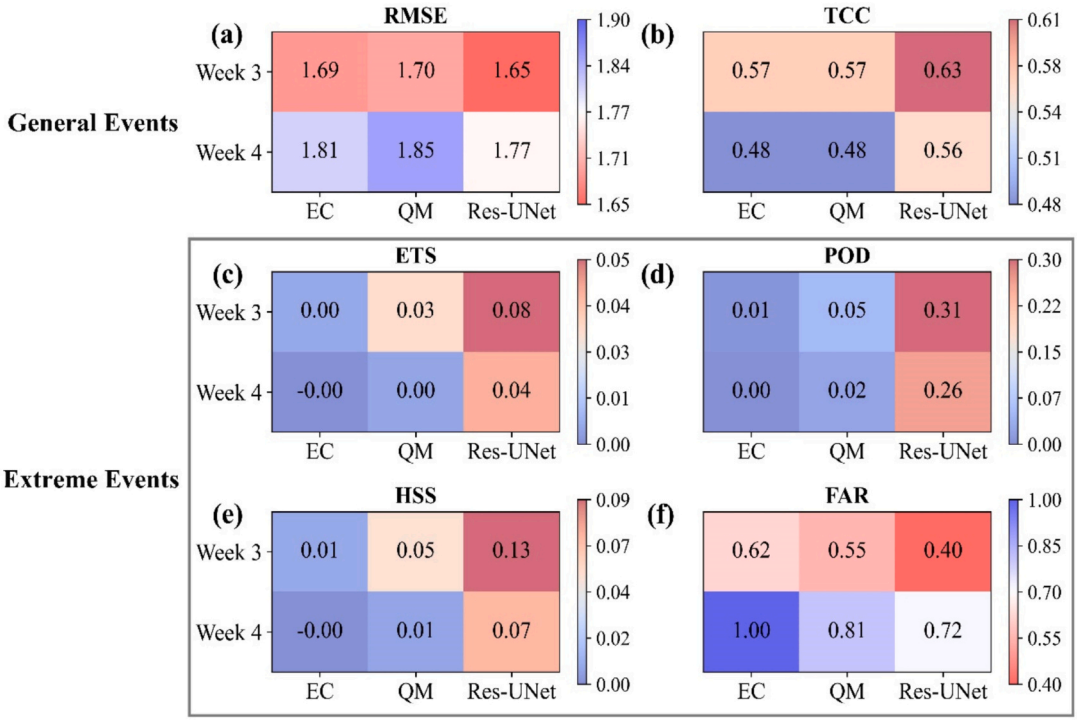

Fig. 2. Evaluation metrics for general statistical events and extreme events from EC, QM and Res-UNet, averaged over the YRB in the testing set (2019–2022) for the 3–4 weeks forecast period. (a-f) represent the RMSE, TCC, ETS, POD, HSS, and FAR, respectively. All metrics are evaluated based on 95 % confidence intervals estimated by bootstrapping with 1000 resamples.

high temperatures over the YRB at 3–4 weeks lead times, we use ETS, POD, HSS, and FAR. Higher ETS, POD, and HSS indicate better performance, while lower FAR reflects improved accuracy. As shown in Fig. 2 (c-f), EC shows lower values for the first three metrics and higher FAR, while QM improves these metrics. However, Res-UNet outperforms both, with the highest ETS, POD, HSS, and lowest FAR. Res-UNet achieves a POD of around 0.3 and reduces FAR by 28 %, demonstrating its superior forecasting capability.

Besides, an exploration of the spatial features of extreme high-temperature forecasting performances for EC, QM, and Res-UNet models is presented, as shown in the POD distributions in Fig. 3. During the 3–4 weeks forecast period, the POD distribution for QM is comparable to that of EC, but it shows slightly better performance. This indicates that traditional QM method has a certain predictive capability for long-term forecasts. Additionally, Res-UNet demonstrates strong forecasting skills, particularly in high-incidence areas of extreme high temperatures, such as Chongqing and Sichuan. Moreover, it outperforms EC and QM in other regions, with its POD values exceeding theirs, extending up to a 4-week lead time. In summary, Res-UNet enhances forecast skills for extreme high temperatures compared to EC and QM for all lead times. Notably, improvements are particularly evident in Chongqing and Sichuan during the 4 weeks lead time, reaching a POD exceeding 0.6.

3.2. Probabilistic forecast

In this study, the output of the Res-UNet is the daily probability from summer sub-seasonal probabilistic prediction. The probabilistic prediction estimates the likelihood of exceeding the high-temperature threshold of 35 °C based on the deterministic forecast. Specifically, the training and testing sets derived from deterministic forecast results are utilized for probabilistic forecast. Fig. S6 presents the spatial distribution of the BSS for probabilistic forecast of extreme high-temperature events for all testing sets from 2019 to 2022 at 3–4 weeks lead times from the EC, QM, and Res-UNet. Compared with EC and QM, the Res-UNet exhibits significantly higher BSS values, indicating superior probabilistic forecast performance at 3–4 weeks lead times (Fig. 4). In addition, the Res-UNet effectively reduces forecast errors, exhibiting lower BS values (Fig. S7). Given the large deterministic forecast errors in 2022 (Fig. S2), the following sections will focus on analyzing the probabilistic performance of the Res-UNet for extreme high temperature, compared to EC and QM during the summer of 2022. The BSS, CRPSS, BS, ROC curve, AUC, and reliability diagram are employed to evaluate the performance of probabilistic forecast.

The CRPSS quantifies the improvement of a probabilistic forecast relative to a reference forecast, with higher values indicating better predictive skill. Fig. S8 further illustrates the spatial distribution of the CRPSS for extreme high temperature over the YRB during 2019–2022, comparing to EC, QM, and the Res-UNet model at 3–4 weeks lead times. Overall, the regional mean CRPSS of Res-UNet is substantially higher than that of EC and QM at both lead times, suggesting that its predicted probability distributions are closer to observations and more accurately capture the occurrence probability of extreme heat events (Fig. 4). By contrast, EC and QM exhibit generally low CRPSS values, with skill concentrated in low-value regions. In particular, QM shows lower skill than EC in some areas at the 4-week lead, indicating limited stability at longer lead times. In contrast, Res-UNet achieves much higher skill over most regions, with CRPSS values exceeding 0.8 in some areas. Overall, Res-UNet significantly outperforms both EC and QM, confirming the superior performance of deep learning methods in subseasonal probabilistic forecasting of extreme high-temperature events.

Figure 5 indicates that the outcomes of the Res-UNet, QM, and EC are the probability of extreme high temperature in the sub-seasonal forecast within the 3–4 weeks forecast period in 2022 over YRB. We observe a significant heatwave occurring from August 3 to August 24, 2022, as illustrated in the upper part of Fig. 5. The heatwave forecasting period includes days with a probability exceeding 50 %, while the overall forecasting probability is the average for this duration (Zhang et al., 2022). Illustrating with a lead time of 3 weeks, the forecast of Res-UNet indicates a 65 % probability of a heatwave occurring between August 3 and August 24, 2022. This suggests a significant chance of extreme temperatures during that period.

In comparison, the probabilities provided by the EC and QM are 45 % and 48 %, respectively. We note that the predicted heatwave period spans from August 14 to August 21 at lead time of 4 weeks. The probabilities for predicting this heatwave are 56 %, 38 %, and 34 % for Res-UNet, QM, and EC, respectively. This indicates that Res-UNet still has an advantage over EC and QM in predicting heatwave occurrences 4 weeks

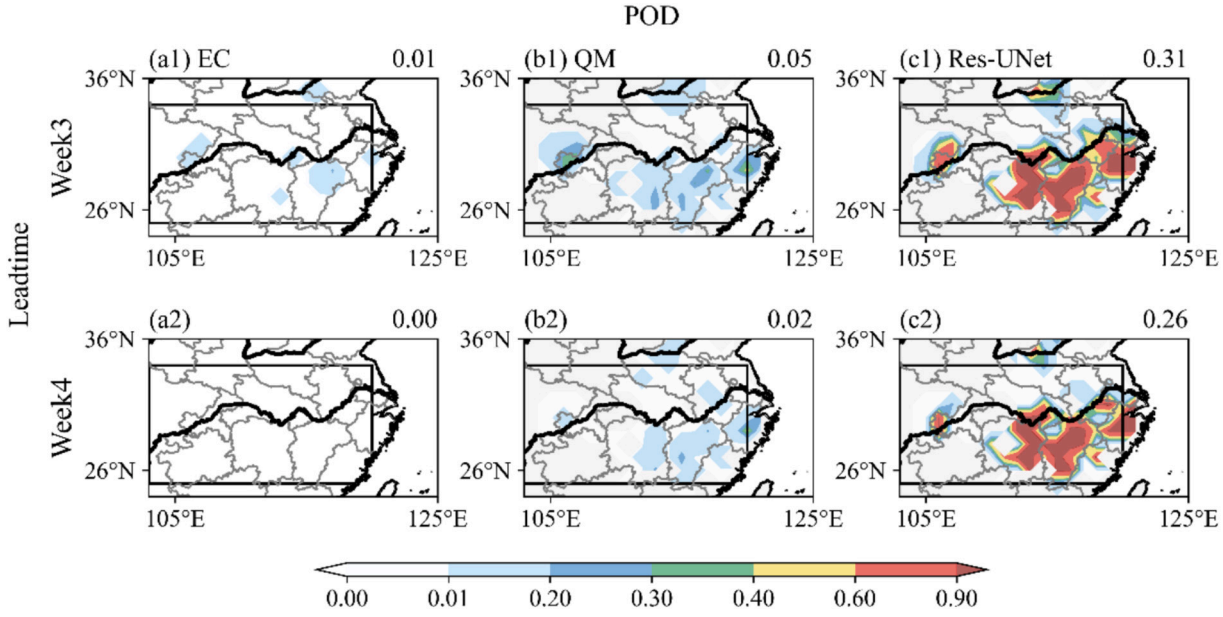

Fig. 3. Spatial distribution of POD for extreme high temperature forecasts for the 3–4 weeks forecast period during 2019–2022. Figures (a1–a2, b1–b2, and c1–c2) indicate EC, QM and Res-UNet, respectively, with the first to second rows corresponding to lead times of 3, 4 weeks. The outlined box represents the study region (YRB: 25°N–34°N, 103°E–120°E). The values in the upper right corner represent the area-mean POD of the black box.

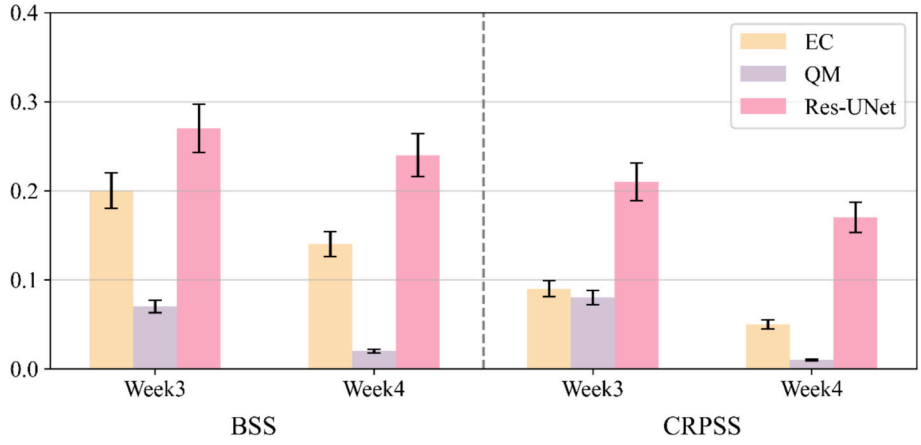

Fig. 4. Brier skill score (BSS) and continuous ranked probability skill score (CRPSS) from EC, QM and Res-UNet for the 3–4 weeks forecast period from 2019 to 2022. All metrics are evaluated based on 95 % confidence intervals estimated by bootstrapping with 1000 resamples.

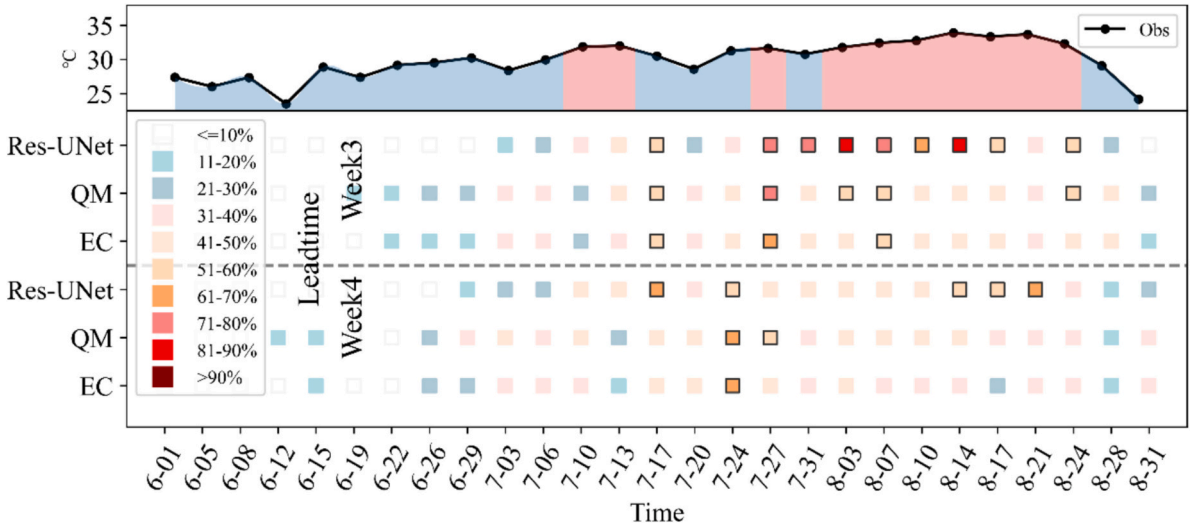

Fig. 5. Observed (upper part; units: °C) and probabilistic forecast of extreme high-temperature (lower part; units: %) by Res-UNet, QM, and EC for the 3–4 weeks forecast period in 2022 over the YRB. Red filled area in the upper part is the observed extreme high-temperature, otherwise blue. The probabilistic levels are represented by different colors in the small square. A black-edged square indicates a probability greater than 50 %. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

ahead of time. In conclusion, the Res-UNet model effectively predicts the heatwave 3 weeks in advance, exhibiting significantly higher confidence. This achievement highlights the model's capability in forecasting extreme climate events. It is noteworthy that due to the bi-weekly nature of forecast days, whether it is for forecasting or observation, the actual start and end dates of heatwaves may vary slightly.

To assess the performance of models on binary classification events (extreme or non-extreme), the ROC curve and AUC are utilized. AUC values closer to 1 indicate better accuracy in distinguishing extremes and non-extreme events. Fig. 6 shows that Res-UNet consistently outperforms EC and QM in predicting extreme high temperatures at 3–4 weeks lead times, with higher AUC values. QM shows slight improvements over EC, but both models exhibit limited predictive strength. To further evaluate the calibration of the probabilistic forecast, reliability diagrams are analyzed. Fig. 7 shows the reliability of probabilistic forecast for Res-UNet, QM, and EC at 3–4 weeks lead times. Res-UNet's reliability curve is closest to the ideal line (gray dashed), demonstrating the best performance and significantly outperforming QM and EC at a 3-week lead time. At the 4-week lead time, all models' performances decline with reliability curves deviating from the ideal line. QM's curve consistently lies above the ideal, indicating conservative forecasts, while EC's curve fluctuates notably with poorer calibration. Although Res-UNet shows slight deviations in the mid-probability range, its curve remains smooth and closer to the ideal line, with the lowest BS, indicating greater robustness in long lead probabilistic forecast. Overall, Res-UNet demonstrates superior probabilistic forecasting ability and skill at 3–4 weeks lead times.

Further analysis of the extreme high-temperature event on August 14, 2022 (Fig. 8) reveals that as lead time increases, all models perform worse. EC fails to capture extreme high temperatures, and QM shows minimal improvement, especially in regions like Chongqing and Sichuan. In contrast, Res-UNet provides more accurate with higher probabilities for these regions, even at a 4-week lead time. Brier Score (BS) analysis confirms Res-UNet's superior performance, with consistently lower BS values, particularly in Chongqing, indicating better accuracy and reliability in forecasting extreme temperatures at sub-seasonal scale (Fig. 8). In a world, Res-UNet reduces deterministic forecast errors, offering robust probabilistic predictions for longer lead times.

The SHAP method is used to enhance the interpretability of machine learning models and analyze the impact of input features on the output of the Res-UNet model. In the decision plot, the gray vertical line in the

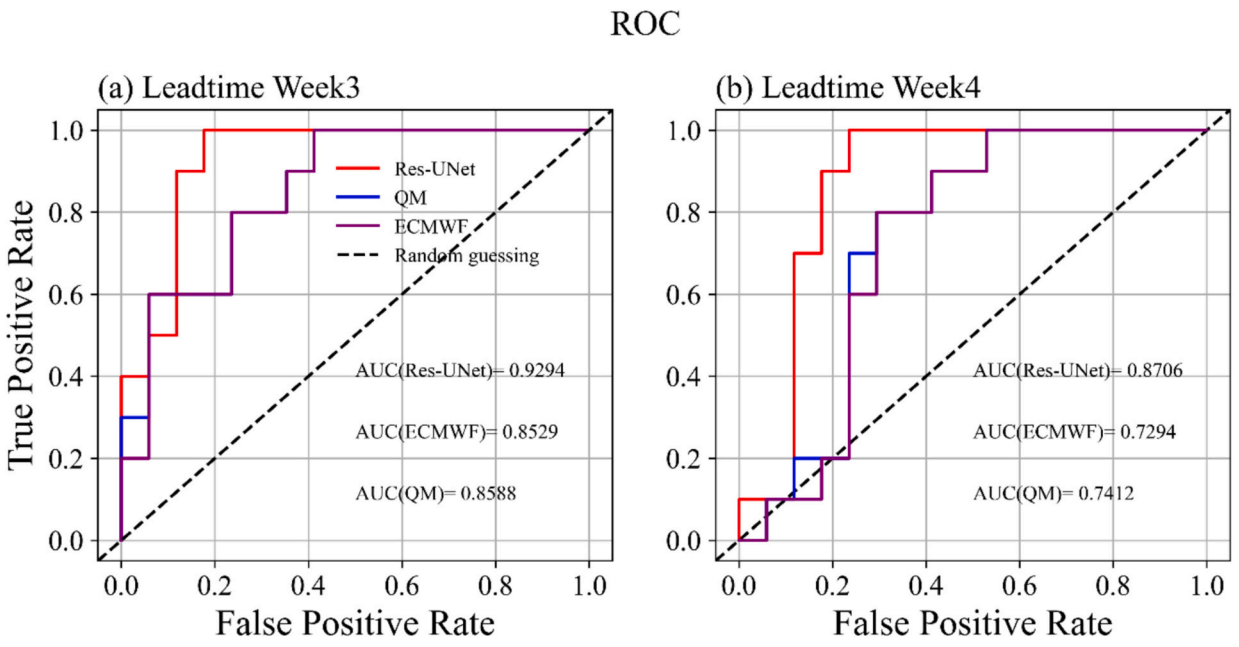

Fig. 6. An illustrating figure of receiver operating characteristic (ROC) curve and area under the curve (AUC) for the 3–4 weeks forecast period in 2022. (a, b) represent 3 and 4 weeks, respectively. Black line represents random guessing.

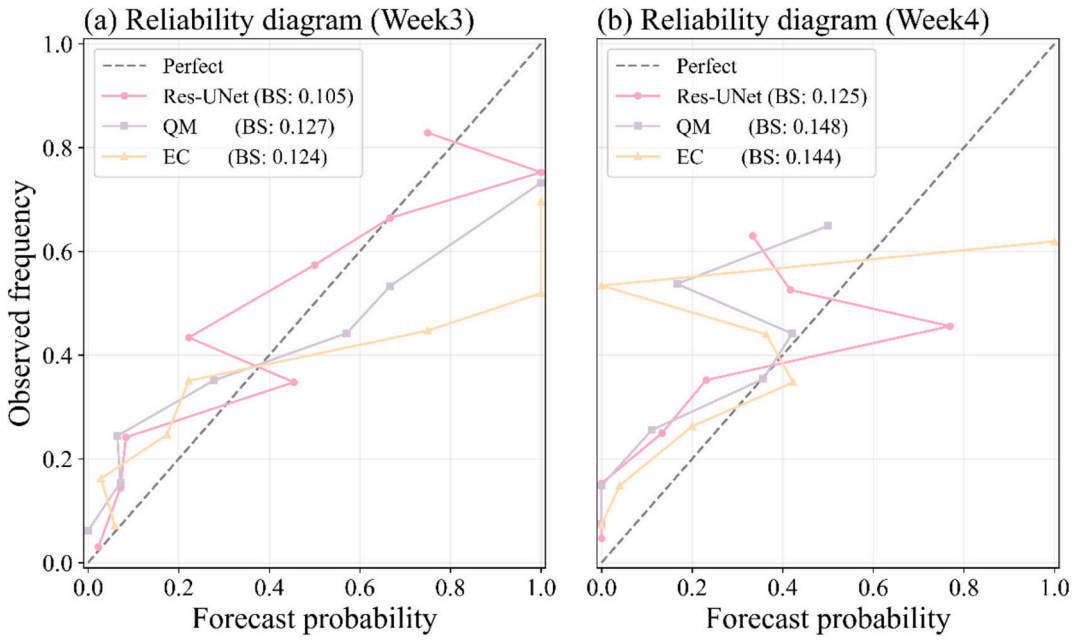

Fig. 7. Reliability diagrams of the probabilistic forecast from Res-UNet (light pink), QM (light purple), and EC (light yellow) at week3 (a) and week4 (b) lead times. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

center marks the model's baseline value, while the colored lines represent the predictions, showing how each feature moves the output value above or below the average predicted value. Feature values are displayed next to the prediction lines for reference. Starting from the bottom of the plot, the prediction lines show how the SHAP values accumulate from the baseline to the model's final score at the top of the plot. Fig. S9 summarizes the SHAP values of feature values and their relationship with changes in the model output for lead times of 3–4 weeks. Results show that tasmax forecast of EC model is the most important factor in the Res-UNet. For a 3-week lead time forecast, the feature values of T500 and Z850 are significant. However, at a 4-week lead time, U200 and Q850 play a crucial role in the Res-UNet. This suggests that incorporating these key factors into the Res-UNet model can enhance its sub-seasonal forecasting capability for extreme high temperatures.

4. Discussions

Based on the results presented in Section 3, we further discuss the strengths, limitations, and potential improvements of the Res-UNet model for sub-seasonal high-temperature forecast. As shown in Fig. S10, the probabilistic forecast indicates that EC and QM generally fail to exceed 50 % probability during the 3–4 weeks forecasting period and are unable to predict more high-temperature areas with higher

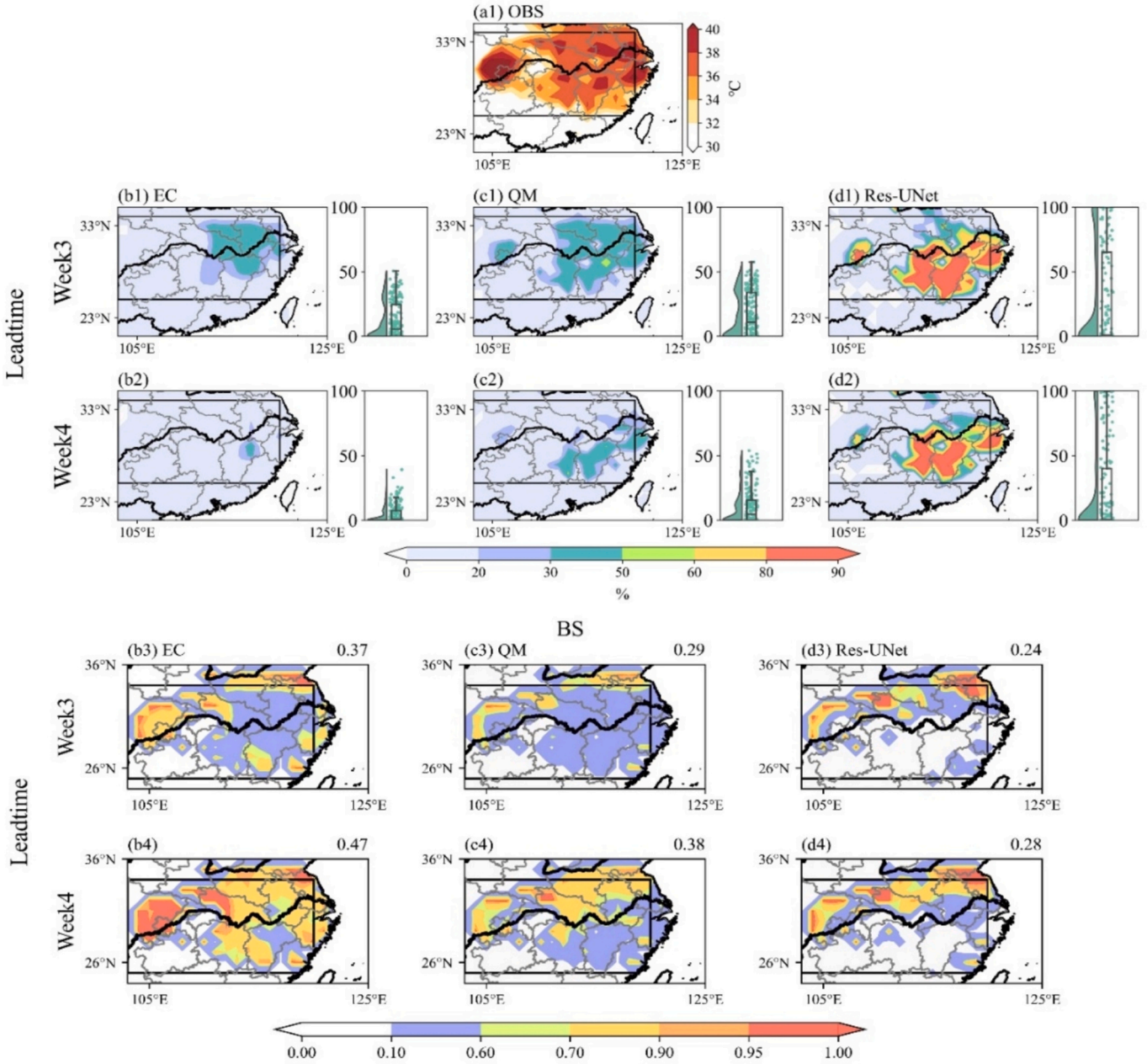

Fig. 8. Spatial distribution of the probabilistic forecast for extreme high temperature and the corresponding Brier Score (BS) on 14 August 2022 over the YRB (black box). The top panels show the probabilistic forecasts of extreme high temperature, with the first column shows observations (a1, units: °C). The second and third rows correspond to probabilistic forecasts (units: %) with lead times of 3 weeks and 4 weeks, respectively. While the bottom panels show the corresponding BS distributions. The three model columns correspond to EC (b1–b4), QM (c1–c4) and Res-UNet (d1–d4). The outlined box denotes the YRB (25°N–34°N, 103°E–120°E). The right-hand side of each subplot in top panels represents the corresponding probability distribution over the YRB. Values in the upper right corner in the bottom panels indicate the area-mean BS over the YRB.

probabilities, highlighting their limited long-term forecasting capability. As lead time increases, their forecasting skill deteriorates. In contrast, Res-UNet aligns well with observed extreme high temperatures: probability values rise when high temperatures occur. It effectively predicts areas experiencing extreme temperatures with higher probability values over the 4-week period, demonstrating its potential for long-range prediction.

In addition, we evaluate the performance of Res-UNet during non-extreme years (2019–2021) and further analyze whether the model exhibits bias toward extreme years in 2022. Specifically, to assess the model's generalization ability under non-extreme conditions, we investigate the TCC of prediction for the 2019–2021 period (Fig is not shown). The result indicates that Res-UNet maintains good predictive performance across most regions, suggesting that it performs well not only during the extreme year of 2022 but also demonstrates strong robustness in non-extreme years, without overfitting to extreme events. Therefore, our model exhibits good stability.

In this study, other machine learning, such as ANN, ConvLSTM, CNN, and RF are used to compare with Res-UNet. From Fig. 9, it is observed that the Res-UNet model performs better than or equal to other models within the 3–4 weeks forecast period. The TCC and RMSE indicate that the Res-UNet exhibits higher skill. In terms of probabilistic forecast, the Res-UNet also has shown better skill, as indicated by its performance metrics. Generally, compared to other models, the Res-UNet shows

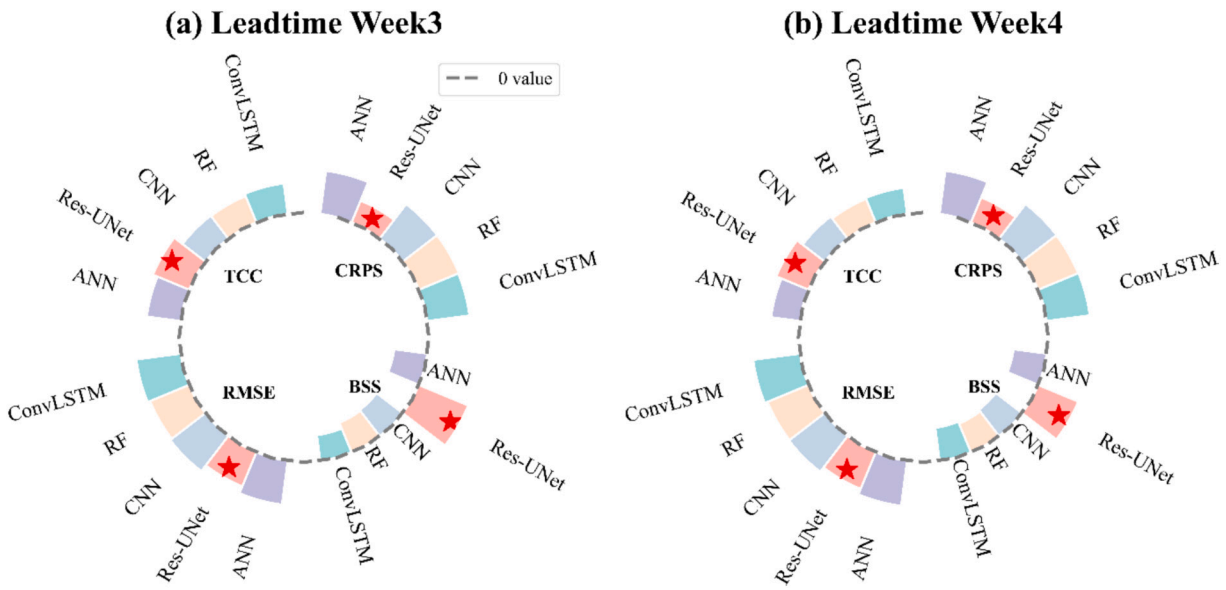

Fig. 9. Comparison of deterministic (TCC, RMSE) and probabilistic forecast (BSS, CRPS) metrics between the Res-UNet, ConvLSTM, RF, CNN, and ANN models. (a) Leadtime Week3, (b) Leadtime Week4.

improvements in predicting extreme high temperatures.

However, the model still exhibits prediction errors in certain marginal regions, particularly on the right side (Fig. S10) during the 3–4 weeks forecast period. This may be attributed to the complexity of climate systems in boundary areas, sparse observational data, and the model's limited ability to handle edge conditions. Future work could improve forecasting performance in these regions by incorporating additional external predictors (such as sea surface temperature and soil moisture) and optimizing the model architecture (e.g., through regional weighting mechanisms or multi-modal inputs).

Notably, this study develops a deep learning model based on Res-UNet, significantly improving extreme high-temperature prediction through a weighted loss function. The model offers strong physical interpretability and robust transferability, allowing it to be adapted via fine-tuning for forecasting various extreme climate events across different regions, demonstrating broad prospects. Additionally, although Res-UNet demonstrates strong capability in identifying high-temperature events in probabilistic forecast, it still struggles to accurately capture the onset and end dates of heatwaves at lead times of 4 weeks or more (Fig. 5). Future research will focus on enhancing the model's temporal sequence modeling capabilities and deepening the analysis of relevant physical processes to improve both forecasting accuracy and interpretability.

5. Conclusions

In this study, we utilize the U-Net framework to establish both deterministic and probabilistic models for sub-seasonal high temperatures over the YRB within the 3–4 weeks lead times. The output derived 2003–2018 is used for training, while the years 2019 to 2022 serve as the testing period to evaluate prediction skills. We compare the performance of Res-UNet with those of EC, QM. The results are as follows:

(1) In deterministic forecast, compared to EC and QM, Res-UNet demonstrates overall advancements and higher improvement, primarily observed in Chongqing and Sichuan, with POD exceeding 0.6 even at a 3–4 weeks lead time.
(2) In terms of probabilistic forecast, Res-UNet outperforms EC and QM at longer lead times, achieving lower BS, higher AUC values, and better calibration and reliability. It also shows improved BSS and CRPSS, indicating more accurate representation of predictive probability distributions and superior ability to capture extreme events. Notably, it shows enhanced skill in probabilistic forecast, especially in Chongqing and Sichuan.
(3) Additionally, Res-UNet provides high-confidence predictions for extreme high temperatures in 2022, accurately forecasting heatwave start and end days up to 3 weeks in advance. Further SHAP analysis reveals that tasmax itself is the most important factor, with T500 and Z850 being significant for a 3-week lead time, while U200 and Q850 become crucial at a 4-week lead time in Res-UNet.

Overall, the **Res-UNet** exhibits superior skill in both deterministic and probabilistic sub-seasonal forecasts of extreme high temperatures. It effectively reduces forecast errors and provides better reliability of extreme heat events. These results highlight the potential of **deep learning methods** to advance deterministic and probabilistic forecasts of extreme climate events at the sub-seasonal scale.

CRediT authorship contribution statement

Shifeng Pan: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Zhicong Yin:** Writing – review & editing, Visualization, Supervision, Resources, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Yi Fan:** Writing – review & editing, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Formal analysis, Data curation. **Tingting Han:** Writing – review & editing, Visualization, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation. **Mingkeng Duan:** Writing – review & editing, Visualization, Validation, Supervision, Project administration, Methodology, Investigation, Funding acquisition. **Huijun Wang:** Writing – review & editing, Visualization, Validation, Supervision, Resources, Investigation, Funding acquisition, Data curation, Conceptualization.

Declaration of competing interest

None.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (Grant No. 2022YFF0801604).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.atmosres.2025.108602.

Data availability

Authors declare that datasets and code for this research are available in the following online repositories. The S2S data can be accessed through the link: https://apps.ecmwf.int/datasets/data/s2s-reforecasts-instantaneous-accum-ecmf/levtype=sfc/type=cf/. The ERA5 reanalysis is obtained from https://cds.climate.copernicus.eu/datasets/reanalysis-era5-single-levels?tab=download.

References

De Bruin, H., Van Den Hurk, B., Kroon, L., 1999. On the temperature-humidity correlation and similarity. Bound.-Layer Meteorol. 93 (3), 453–468.

Ding, S., Zhi, X., Lyu, Y., Ji, Y., Guo, W., 2024. Deep learning for daily 2-m temperature downscaling. Earth. Space. Sci. 11. https://doi.org/10.1029/2023EA003227 e2023EA003227.

Edward, N.L., 1963. Deterministic nonperiodic flow. J. Atmos. Sci. 20 (2), 130–141. https://doi.org/10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2.

Fan, K., Wang, H.J., Choi, Y.J., 2008. A physically-based statistical forecast model for the middle-lower reaches of the Yangtze River Valley summer rainfall. Chin. Sci. Bull. 53, 602–609. https://doi.org/10.1007/s11434-008-0083-1.

Gneiting, T., Katzfuss, M., 2014. Probabilistic forecasting. Ann. Rev. Stat. Appl. 1, 125–151. https://doi.org/10.1146/annurevstatistics-062713-085831.

Gneiting, T., Raftery, A.E., Westveld, A.H., Goldman, T., 2005. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. Mon. Weather Rev. 133 (5), 1098–1118.

Guo, Q., Liu, X.W., Wu, T.W., et al., 2017. Verifcation and correction of East China summer rainfall prediction based on BCC_CSM. Chin. J. Atmos. Sci. 41 (1), 71–90. https://doi.org/10.3878/j.issn.1006- 9895.1602.15280 (in Chinese).

Hamill, T.M., Colucci, S.J., 1997. Verification of Eta-RSM short-range ensemble forecasts. Mon. Weather Rev. 125 (6), 1312–1327.

He, S., Li, X., DelSole, T., Ravikumar, P., Banerjee, A., 2021. Sub-Seasonal climate forecasting via machine learning: challenges, analysis, and advances. Proceed. AAAI Confer. Artif. Intell. 35 (1), 169–177. https://doi.org/10.48550/arXiv.2006.07972s.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., et al., 2020. The ERA5 global reanalysis [Dataset]. Q. J. R. Meteorol. Soc. 146 (730), 1999–2049. https://doi.org/10.1002/qj.3803.

Hsu, P.C., Lee, J.Y., Ha, K.J., Tsou, C.H., 2017. Influences of boreal summer intraseasonal oscillation on heat waves in monsoon Asia. J. Clim. 30 (18), 7191–7211. https://doi.org/10.1175/JCLI-D-16-0505.1.

Hsu, P.C., Zang, Y.X., Zhu, Z.W., Li, T., 2020. Subseasonal-to-seasonal(S2S) prediction using the spatial-temporal projection model (STPM). Trans. Atmos. Sci. 43 (1), 212–224. https://doi.org/10.13878/j.cnki.dqkxxb.2019102800 (in Chinese).

Hu, Y.F., Yin, F.K., Zhang, W.M., 2021. Deep learning-based precipitation bias correction approach for Yin–He global spectral model. Meteorol. Appl. 28 (5), e2032. https://doi.org/10.1002/met.2032.

Hua, W., Dai, A., Qin, M., Hu, Y., Cui, Y., 2023. How unexpected was the 2022 summertime heat extremes in the middle reaches of the Yangtze River? Geophys. Res. Lett. 50 (16), e2023GL104269. https://doi.org/10.1029/2023GL104269.

Jin, W., Zhang, W., Hu, J., Weng, B., Huang, T., Chen, J., 2022. using the residual network module to correct the sub-seasonal high temperature forecast. Front. Earth Sci. 9, 760766. https://doi.org/10.3389/feart.2021.760766.

Kolachian, R., Saghafian, B., 2019. Deterministic and probabilistic evaluation of raw and post-processed sub-seasonal to seasonal precipitation forecasts in different precipitation regimes. Theor. Appl. Climatol. 137, 1479–1493.

Lei, L., Hsu, P.C., Gao, Q.J., Xie, J.H., 2022. Extended-range forecasting method of summer daily maximum temperature in the Yangtze River Basin based on convolutional neural network. Trans. Atmos. Sci. 45 (6), 835–849. https://doi.org/10.13878/j.cnki.dqkxxb.20211101001.

Li, Q.X., Huang, J.Y., 2011. Threshold values on extreme high temperature events in China. J. Appl. Meteorol. Sci. 22 (2), 138–144 (in Chinese).

Li, R.X., Sun, J.Q., 2018. Interdecadal variability of the large-scale extreme hot event frequency over the middle and lower reaches of the Yangtze River basin and its related atmospheric patterns. Atmos. Oceanic Sci. Lett. 11 (1), 63–70. https://doi.org/10.1080/16742834.2018.1410057.

Li, W., Pan, B., Xia, J., Duan, Q., 2022. Convolutional neural network-based statistical post-processing of ensemble precipitation forecasts. J. Hydrol. 605, 127301. https://doi.org/10.1016/j.jhydrol.2021.127301.

Liang, P., Lin, H., 2018. Sub-seasonal prediction over East Asia during boreal summer using the ECCC monthly forecasting system. Clim. Dyn. 50, 1007–1022. https://doi.org/10.1007/s00382-017-3658-1.

Lin, H., Mo, R., Vitart, F., 2022. The 2021 western north American heatwave and its subseasonal predictions. Geophys. Res. Lett. 49, e2021GL097036. https://doi.org/10.1029/2021GL097036.

Lorenc, A.C., 1986. Analysis methods for numerical weather prediction. Q. J. R. Meteorol. Soc. 112 (474), 1177–1194. https://doi.org/10.1002/qj.49711247414.

Lu, R., Xu, K., Chen, R., Chen, W., Li, F., Lv, C., 2023. Heat waves in summer 2022 and increasing concern regarding heat waves in general. Atmos. Oceanic Sci. Lett. 16 (1), 100290. https://doi.org/10.1016/j.aosl.2022.100290.

Lundberg, S., Lee, S.-I., 2017. A Unified Approach to Interpreting Model Predictions. https://doi.org/10.48550/arXiv.1705.07874.

Lyu, Y., Zhu, S., Zhi, X., Ji, Y., Fan, Y., Dong, F., 2023. Improving subseasonal-to-seasonal prediction of summer extreme precipitation over southern China based on a deep learning method. Geophys. Res. Lett. 50, e2023GL106245. https://doi.org/10.1029/2023GL106245.

Ma, F., Ye, A., Deng, X., Xu, J., Miao, C., 2016. Evaluating the skill of NMME seasonal precipitation ensemble predictions for 17 hydroclimatic regions in continental China. Int. J. Climatol. 36, 132–144.

Pan, B., Anderson, G.J., Goncalves, A., Lucas, D.D., Bonfils, C.J.W., Lee, J., 2022. Improving seasonal forecast using probabilistic deep learning. J. Adv. Model. Earth Syst. 14, e2021MS002766. https://doi.org/10.1029/2021MS002766.

Pan, S.F., Yin, Z.C., Fan, Y., Duan, M.K., 2024. A Method for Subseasonal Extreme high-temperature Prediction Based on an Improved U-Net Network. Patent No. 202410798747X. 202407 (in Chinese).

Pegion, K., Kirtman, B.P., Becker, E., Collins, D.C., Kim, H., 2019. The Subseasonal Experiment: a multimodel subseasonal prediction experiment. Bull. Am. Meteorol. Soc. 100 (10), 2043–2060. https://doi.org/10.1175/BAMS-D-18-0270.1.

Pyrina, M., Domeisen, D.I.V., 2023. Subseasonal predictability of onset, duration, and intensity of European heat extremes. Q. J. R. Meteorol. Soc. 149, 84–101. https://doi.org/10.1002/qj.4394.

Raftery, A.E., Gneiting, T., Balabdaoui, F., Polakowski, M., 2005. Using Bayesian model averaging to calibrate forecast ensembles. Mon. Weather Rev. 133, 1155–1174.

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.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation[C]//. Intern. Conf. Med. Image Comput. Comp.-assisted Intervent. 9351, 234–241. https://doi.org/10.1007/978-3-319-24574-4_28.

Sloughter, M., Raftery, A.E., Gneiting, T., Fraley, C., 2007. Probabilistic quantitative precipitation forecasting using Bayesian model averaging. Mon. Weather Rev. 135, 3209–3220.

Sun, B., Wang, H., Huang, Y., Yin, Z., Zhou, B., Duan, M., 2023. Characteristics and causes of the hot-dry climate anomalies in China during summer of 2022. Trans. Atmos. Sci. 46 (1), 1–8. https://doi.org/10.13878/j.cnki.dqkxxb.20220916003. (in Chinese).

Sun, G.W., Xin, F., Kong, C.Y., Chen, B.M., Jin, H.E., 2010. Atmospheric low-frequency oscillation and extended range forecast. Plateau Meteorol. 29 (5), 1142–1147. https://doi.org/10.3788/gzxb20103906.0998.

Tan, J., Zheng, Y., Song, G., Kalkstein, L.S., Kalkstein, A.J., Tang, X., 2007. Heat wave impacts on mortality in Shanghai, 1998 and 2003. Int. J. Biometeorol. 51 (3), 193–200. https://doi.org/10.1007/s00484-006-0058-3.

Tang, S.K., Qiao, S.B., Feng, T.S., Fan, P.Y., Liu, J.Y., Zhao, J.H., Feng, G.L., 2023. Predictability of the unprecedented 2022 late summer Yangtze River Valley and Tibetan Plateau heatwaves by the NCEP CFSv2. Atmos. Res. 296, 107053. https://doi.org/10.1016/j.atmosres.2023.107053.

Tong, Y., Gao, X., Han, Z., Xu, Y., Xu, Y., Giorgi, F., 2021. Bias correction of temperature and precipitation over China for RCM simulations using the QM and QDM methods. Clim. Dyn. 57 (5–6), 1425–1443. https://doi.org/10.1007/s00382-020-05447-4.

Toth, Z., 2001. Ensemble forecasting in WRF. Bull. Am. Meteorol. Soc. 82, 695–697.

Vigaud, N., Tippett, M.K., Robertson, A.W., 2018. Probabilistic skill of subseasonal precipitation forecasts for the East Africa–West Asia sector during September–May. Weather Forecast. 33, 1513–1532.

Vitart, F., 2017. Madden–Julian Oscillation prediction and teleconnections in the S2S database. Q. J. R. Meteorol. Soc. 143 (706), 2210–2220. https://doi.org/10.1002/qj.3079.

Vitart, F., Ardilouze, C., Bonet, A., Brookshaw, A., Chen, M., Codorean, C., et al., 2017. The subseasonal to seasonal (S2S) prediction project database. Bull. Am. Meteorol. Soc. 98 (1), 163–173. https://doi.org/10.1175/BAMS-D-16-0017.1.

Vitart, F., Robertson, A.W., Spring, A., Pinault, F., Roškar, R., Cao, W., et al., 2022. Outcomes of the WMO prize challenge to improve subseasonal to seasonal predictions using artificial intelligence. Bull. Am. Meteorol. Soc. 103 (12), E2878–E2886. https://doi.org/10.1175/BAMS-D-22-0046.1.

Vitart, F., et al., 2025. The WWRP/WCRP S2S project and its achievements. Bull. Am. Meteorol. Soc. 106, E791–E808. https://doi.org/10.1175/BAMS-D-24-0047.

Weyn, J.A., Durran, D.R., Caruana, R., Cresswell-Clay, N., 2021. Sub-seasonal forecasting with a large ensemble of deep-learning weather prediction models. J. Adv. Model. Earth Syst. 13, e2021MS002502. https://doi.org/10.1029/2021MS002502.

White, C.J., Carlsen, H., Robertson, A., 2017. Potential applications of subseasonal-to-seasonal (S2S) predictions. Meteorog. Atmos. Phys. 24 (3), 315–325. https://doi.org/10.1002/met.1654.

Wilks, D.S., 2013. The calibration simplex: a generalization of the reliability diagram for three-category probability forecasts. Weather Forecast. 28, 1210–1218. https://doi.org/10.1175/WAF-D-13-00027.1.

Xie, J., Yu, J., Chen, H., Hsu, P.C., 2020. Sources of Subseasonal Prediction Skill for Heatwaves over the Yangtze River Basin Revealed from three S2S Models. Adv. Atmos. Sci. 37 (12), 1435–1450. https://doi.org/10.1007/s00376-020-0144-1.

Xie, J., Hsu, P.-C., Hu, Y., Zhang, H., Ye, M., 2024. Advancing subseasonal surface air temperature and heat wave prediction skill in China by incorporating scale interaction in a deep learning model. Geophys. Res. Lett. 51, e2024GL111076. https://doi.org/10.1029/2024GL111076.

Yin, Z., Zhou, B., Duan, M., Chen, H., Wang, H., 2023. Climate extremes become increasingly fierce in China. Innovation 4 (2), 100406. https://doi.org/10.1016/j.xinn.2023.100406.

Yin, Z.C., Song, X.L., Zhou, B.T., Jiang, W.H., Chen, H.P., Wang, H.J., 2024. Traditional Meiyu-Baiu has been suspended by global warming. Natl. Sci. Rev. 11 (7), nwae166. https://doi.org/10.1093/nsr/nwae166.

Zhang, L., Yang, T., Gao, S., Hong, Y., Zhang, Q., Wen, X., Cheng, C., 2023. Improving subseasonal-to-seasonal forecasts in predicting the occurrence of extreme precipitation events over the contiguous U.S. using machine learning models. Atmos. Res. 281, 106502. https://doi.org/10.1016/j.atmosres.2022.106502.

Zhang, W., Gao, J., Lai, Q., Chi, Y., Su, T., 2022. Probabilistic Forecast of the Extended Range Heatwave over Eastern China. Front. Earth Sci. 9, 810579. https://doi.org/10.3389/feart.2021.810579.

Zhou, J., Liu, F., 2025. Improving subseasonal forecasting of East Asian monsoon precipitation with deep learning. Atmos. Oceanic Sci. Lett., 100520 https://doi.org/10.1016/j.aosl.2024.100520.

Zhu, S., Zhang, L., Jiang, H., Lyu, Y., Fan, Y., Guo, Z., et al., 2023. Pattern projection calibrations on subseasonal forecasts of surface air temperature over East Asia. Weather Forecast. 38 (6), 865–878. https://doi.org/10.1175/WAF-D-22-0046.1.

Zhu, T., Flavio, F.D., Ive, D.S., 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.

Zhu, Y., Zhi, X., Lyu, Y., Zhu, S., Tong, H., Ali, M., et al., 2022. Forecast calibrations of surface air temperature over Xinjiang based on U-net neural network. Front. Environ. Sci. 10, 1011321. https://doi.org/10.3389/fenvs.2022.1011321.