Preprints.org

Article

Hybrid Post-processing on GEFSv12 Reforecast for Summer Maximum Temperature Ensemble Forecasts on Extended Range Time Scale over Taiwan

Murali Nageswara Rao Malasala [*] , Yuejian Zhu , Vijay Tallapragada , Meng-Shih Chen

Original Research Article

Hybrid Post-Processing on GEFSv12 Reforecast for Summer Maximum Temperature Ensemble Forecasts on Extended Range Time Scale over Taiwan

M. M. Nageswararao [1,*], Yuejian Zhu [2], Vijay Tallapragada [2] and Meng-Shih Chen [3]

[1] CPAESS, University Corporation for Atmospheric Research at NOAA/NWS/NCEP/EMC, College Park, MD 20740; murali.n.malasala@noaa.gov

[2] NOAA/NWS/NCEP/EMC, College Park, MD 20740; yuejian.zhu@gmail.com (Y.Z.); vijay.tallapragada@noaa.gov (V.T.)

[3] Central Weather Bureau, Taipei, Taiwan; mschen@mfc.cwb.gov.tw

[*] Correspondence: muralinagesh.ocean@gmail.com, murali.n.malasala@noaa.gov; Tel.: (+14698655286)

Abstract: Taiwan is highly susceptible to global warming, experiencing a 1.4°C increase in air temperatures from 1911-2005, which is twice the average for the Northern Hemisphere. This has led to higher rates of respiratory and cardiovascular mortality. Accurately predicting maximum temperatures during the summer season is crucial, but numerical weather models become less accurate and more uncertain beyond five days. To improve forecast reliability, statistical post-processing is needed to address systematic errors. In September 2020, NOAA NCEP implemented the Global Ensemble Forecast System version 12 (GEFSv12) to help manage climate risks. This study developed a Hybrid statistical post-processing method that combines Artificial Neural Networks (ANN) and Quantile mapping (QQ) approaches to predict daily maximum temperatures and extremes in Taiwan during the summer season. The Hybrid technique, utilizing deep learning techniques, was applied to the GEFSv12 reforecast data and evaluated against ERA5 reanalysis. The Hybrid technique was the most effective among the three techniques tested. It had the lowest bias, RMSE, and highest correlation coefficient. It successfully reduced the warm bias and overestimation of Tmax extreme days. This led to improved prediction skills for all forecast lead times. Compared to ANN and QQ, the Hybrid method was more effective in predicting summer daily Tmax and its extremes on an extended-range time scale deterministic and ensemble probabilistic forecasts over Taiwan.

Keywords: deep learning; Ensemble Forecast; GEFSv12; extended range time scale; Hybrid Post-processing; maximum temperature; Taiwan

1. Introduction

Temperature, one of the weather components, is the measurement of how hot or cold the environment is and significantly impacts activities of sectors such as the energy industry, aviation industry, communication pollution dispersal, and agriculture. Climate change has become this century's most severe scientific and social challenge. The IPCC report [1] shows that in the last 50 years, the annual mean temperature has had an increasing linear trend of about 0.13 °C per decade, almost doubled compared to the past century. In recent decades, extremely high temperatures, such as heat waves, have been recurrently observed across the globe, such as in Asia [2–5], USA [6], Europe [7]. Under the rising trend of global warming, heat waves are projected to augment in frequency, intensity, duration, and spatial extent [1,8]. Heatwaves are one of the dangerous natural hazards that cause increased deaths and emergency hospital admissions worldwide, particularly impacting older people, children, and patients with chronic diseases [2,9].   The US National Weather Service evident that the annual casualties due to heat waves were more than   many dynamic natural disasters such as floods, hurricanes, lightning, and tornadoes (http://www.nws.noaa.gov/om/hazstats.shtml).

Taiwan is one of East Asia's subtropical islands, making it susceptible to the extreme weather and climatic changes brought on by global warming [10]. According to the IPCC assessment [11], the rise in the air temperature in Taiwan (1.4 °C) from 1911 to 2005 is near twice the increase (0.7 °C) in the North Hemisphere. It is evident that in Taipei during 1994–2003, each 1°C increase in surface air temperature above 31.5 °C increases respiratory mortality by approximately 9.3% (range, 4.1–14.8%), whereas each 1°C increase above 25.2 °C increases cardiovascular mortality by around 1.1% (range, 0.3–1.9%) [12].

The extended-range forecasts are generally used to predict weather and climate extremes such as heatwaves, cold waves, droughts, and floods. These forecasts can provide relevant weather information, such as the timing of the onset of a rainy season, the risk of extreme rainfall events, heat waves, etc. However, still, there is a well-known gap in current numerical prediction systems for extended range time scale. This gap falls between medium-range weather forecasts (up to 10 days) and seasonal climate predictions (longer than one month). The initial conditions of the atmosphere influence medium-range weather forecasts. In contrast, predictions of the seasonal climate are more affected by slowly evolving surface boundary conditions, such as the sea surface temperature and soil moisture content [13]. Predictions on the extended range timescale have progressed in some regions and seasons [14], despite the full potential of their predictability requiring further exploration.

In past years, there has been a significant improvement in the accuracy of short and medium-range weather forecasts around the world, particularly in extra-tropical regions, which benefits from advanced numerical modeling. However, the same cannot be said for tropical area, such as monsoon regions, where prediction skill remains inadequate [15–17]. This is mainly due to the complexity of tropical processes, which are influenced by interactions between ocean-land-atmosphere, atmospheric circulation, convection, cloud and radiation, precipitation and moisture on different spatial and temporal scales. To improve prediction skill on this time scale in small regions such as Taiwan Island, global models need to be improved to better represent land-sea contrast and topography [18]. In addition to improving the global models, post-processing techniques are also essential for improving extended-range forecasts in small regions. Post-processing techniques can be used to correct systematic errors in the model output, such as bias in the mean and variance of the forecast variables. Numerous studies [19–21] have shown that the post-processing of GCMs raw forecasts is essential. Various post-processing approaches with varying complexities and statistical basis have been developed to calibrate raw GCM forecasts [22,23]. Computationally efficient approaches, such as the rank histogram calibration method [20], "poor man's ensemble" [24], analog method [25], frequency match method (FMM; [26]), and the quantile mapping method ([16,17,27] have been proposed due to their ease of implementation and low computation cost. However, these methods may not be reliable and skillfully calibrated forecasts [28]. Data-driven models, such as machine-learning models, are increasingly used in post-processing. Comparisons of these models show that their performance varies with study areas, GCMs, and evaluation metrics [23]. Therefore, there is no single best post-processing model [29,30]. To provide reliable and skillful forecasts, an effective post-processing approach is needed that is unbiased, reliable in ensemble spread, and at least as good as the climatology reference forecasts.

The NOAA NCEP has implemented the Global Ensemble Forecast System version 12 (GEFSv12) to support stakeholders for sub-seasonal forecasts, hydrological, and other meteorological applications [27,31–33]. This model provides consistent reforecast products for the period 2000-2019, which are available on Amazon Web Services (AWS, https://registry.opendata.aws/noaa-gefs/) and accessible to the public. In this study, the Artificial Neural Network combined with Quantile mapping (ANN-QQ; hereafter termed as Hybrid Post-processing) based on statistical post-processing technique is applied to NCEP GEFSv12 reforecast raw products for predicting summer (June through September; JJAS) surface air maximum temperature (Tmax) and associated extremes (Tmax ≥ 90th percentile of annual Tmax) on   an extended-range time scale over Taiwan. The paper is organized as follows: a brief description of the data and analysis methodologies is given in Section. 2. The results are discussed in Section. 3, and the broad conclusions are presented in Section 4.

2. Data and Methodology

2.1. Data used:

The surface air maximum temperature (Tmax) products of the NCEP GEFSv12 over Taiwan island for reforecast period (2000-2019) has been obtained from Amazon AWS. These products are generated from daily 00 UTC initial conditions, out to 16-day lead-time forecasts for 5 ensemble members. The reforecast products are extended up to day-35 with weekly once (every Wednesday) 00 UTC initial conditions for 11 ensemble members [27]. GEFSv12 reforecast products based on the Global Forecast System version 15.1 (GFSv15.1). It uses the FV3 Cubed-Sphere dynamical core [34] with a horizontal resolution of ~25 km (C384 grid) and 64 hybrid vertical levels, with the top layer centered at 0.27 hPa (~55 km). A modified scale-aware convection parameterization scheme is included in GEFSv12 model physics to reduce the excessive cloud-top cooling for model stabilization [35]. The hybrid Eddy-Diffusivity Mass-flux (EDMF) scheme is used to simulate vertical mixing in the planetary boundary layer [36], while the GFDL-based cloud microphysics scheme predicts five cloud species [31,32]. To estimate the shortwave and longwave radiative fluxes, the rapid radiative transfer model (RRTM), which was developed at Atmospheric and Environmental Research, is used [37]. Chun and Baik (1998) developed a scheme for convective gravity wave drag [38], while the GFS orographic gravity wave drag and mountain blocking schemes are based on Alpert's (1988) study [39]. A two-tiered approach is used to derive the SST boundary conditions, which accounts for the day-to-day variability of Sea Surface Temperature (SST) and Near Sea Surface Temperature (NSST) respectively [40–42]. The GEFSv12 forecast system uses SKEB [43,44] and SPPTs [45,46] to represent model uncertainty. Further details on the configuration and impacts of the individual components can be found in Zhou [31,32].

In this study, the GEFSv12 Tmax reforecast products have been used, which are based on every day 00 UTC initial conditions for day-1 to 16 lead time forecasts with 5 members. The reforecast data is available in grib2 format at 3-hour intervals at 0.25° resolution for the first 10 days forecasts while 6-hour intervals at 0.5° resolution beyond 10 days forecasts. For uniformity, day-1 to 10 forecasts are also considered the same horizontal resolution as day-11 to 16 forecasts. ECMWF Reanalysis version 5 (ERA5) maximum 2m air temperature (Tmax) over Taiwan for the same period (2000-2019) (https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels?tab=form) has been used as a reference for the performance evaluation of GEFSv12 for summer Tmax and associated Tmax extreme events (daily Tmax >90th percentile) over Taiwan with day-1 to 16 forecast lead times [47].

2.2. Calibration Methods:

2.2.1. Quantile Mapping:

A suitable statistical post-processing technique is highly required to calibrate any GCM raw forecast products based on the reforecast period uncertainty for skillful forecast guidance and to increase its usability. In this study, the quantile mapping post-processing technique is used as one of the calibration methods to consider as a benchmark to evaluate the Artificial Neural Network (ANN) and the ANN combination with Quantile mapping (ANN-QQ, hereafter mentioned as Hybrid). These calibration methods on NOAA GEFSv12 reforecast products for deterministic and ensemble probabilistic forecasts of summer Tmax and its extremes on an extended range time scale over Taiwan with Day-1 to 16 forecast lead times have been evaluated against ERA5 reanalysis.

The QQ method, also known as histogram equalization or rank matching [16,17,48], is used to transform model data into bias-corrected data statistically to increase the usability of model products after calibration. The daily Tmax statistics for ERA5 reanalysis and GEFSv12 reforecasts were determined separately for each lead time (Day-1 to Day-16) and grid point of Taiwan. This calibration method is applied independently to each of the five ensemble members and each forecast lead time. To increase the sample size, a 31-day moving window is used, with the forecast day as the center. This results in a sample size of 620-time steps (31 days * 20 years) for each day and each lead time

forecast at a grid point. For the gridpoint on 01st June 2000 with a day-1 forecast lead time, the sample of daily Tmax from 17th May to 16th June from ERA5 and GEFSv12 reforecast period was used. For each day from June 1st to September 30th, the same procedure is implemented separately for each lead time and ensemble member at a gridpoint to approximate the daily Tmax intensity distributions from ERA5 and GEFSv12 reforecasts. This technique utilizes the empirical probability distributions of ERA5 and GEFSv12 Tmax values to generate a calibrated output.   The bias-corrected value for the Tmax forecast of the GEFSv12 model   (Q) can be calculated by taking the inverse of the cumulative distribution function (CDF) of ERA5 values (CDF_{ERA5}^{-1}) at the probability corresponding to the raw GEFSv12 output CDF ($CDF_{GEFSv12}$) at a particular value (F_t).

$$Q = CDF_{ERA5}^{-1}\big(CDF_{GEFSv12}(F_t)\big)$$

The technique of quantile mapping is a transformation between CDFs of the ERA5 and GEFSv12 models. The Leave-one-out Cross-Validation (LOOCV) procedure is implemented. The raw and QQ-calibrated forecasts are referred to as Raw-GEFSv12 and QQ-GEFSv12, respectively.

2.2.2. Artificial Neural Network (ANN):

The IPCC report [11] has highlighted the significant global temperature fluctuations, making it increasingly difficult for meteorological scientists to predict temperature under climate change. Temperature is an atmospheric variable that changes continuously over time, forming a non-linear time series. This complexity and non-linearity make extended-range prediction of temperatures a highly complex phenomenon. One of the study [49] suggest that an ANN based post-processing technique can be used to more accurately to predict temperature than the conventional technique, as the ANN is better equipped to handle the complex non-linear physical variables of the atmosphere. This non-linear technique can help to identify irregularities in the current system. The ANN can model non-linear atmospheric variables without the need for making assumptions, which is more advantageous than the statistical approaches used for temperature predictions that often involve making assumptions about the data, such as assuming that the data follows a normal distribution or that the data is stationary. In addition, they may assume that the connections between variables are linear, which may not always be accurate. Making assumptions in temperature prediction can sometimes lead to more confusion than the expected results [50]. The ANN is a powerful tool for quickly processing data, based on the concept of parallel, machine-based processing inspired by biological neurons. This system of interconnected neurons works together to solve complex problems, resulting in more accurate temperature predictions than those generated by traditional statistical methods [50]. The adaptive nature of ANNs is one of its most important features, as it allows the weights of the model to be updated in order to learn the relationship between inputs and outputs directly from the existing data. This feature enables ANNs to reduce long training times and the need for large amounts of data compared to other statistical techniques [51]. Furthermore, this feature allows ANNs to create a relationship between inputs and outputs, which can be used to predict outputs [52]. The ANN is a data-driven approach that can identify non-linear relationships between input and output parameters without the need to solve complex partial differential equations [53]. This makes it an ideal tool for temperature predictions, which has been demonstrated in numerous studies involving atmospheric time series data [54]. This study implements an ANN calibration method on GEFSv12 reforecast products for summer Tmax and associated Tmax extremes over Taiwan for all forecast lead times (Day-1 to 16).

ANNs can be single-layer or multi-layer, with the single-layer feed-forward neural network model having each input unit connected to an output unit without any connections between the input units, while the multilayer feed-forward neural network has its input and output units linked [55,56]. Back Propagation Neural Network (BPNN) is a popular ANN architecture that uses a gradient descent algorithm to minimize the difference between the actual output and the desired target [57,58]. This type of ANN is constructed, evaluated, and tested for accuracy. A study conducted by Hecht-Nielsen [59] to determine the optimal number of hidden neurons in a single hidden layer of neurons to prevent overfitting. To do this, they used double-cross-validation. However, the performance of

an ANN is highly dependent on the data used for training and the parameters used, such as initial weights, learning rate, momentum, epoch, and activation functions. Therefore, the structure of the ANN must be tailored to the characteristics of the datasets in order to achieve optimal performance. In this study, pre-processing techniques were used to identify and eliminate outliers, reduce noise, and normalize the range of inputs and outputs for the daily maximum temperature from the 5 ensemble members of GEFSv12. The Neural Network Toolbox in MATLAB was used to train, visualize, and simulate ANNs. To determine the optimal ANN structure for a particular day forecast, a double cross-validation procedure was employed. This involved leaving out one data from the 620-sample data and fitting the ANN model to the remaining data in a cross-validation mode. The performance of each iteration was monitored using metrics such as Mean Square Error (MSE) and Root Mean Squared Error (RMSE) while the number of hidden neurons was increased from 1 to 20.

Some of the Research studies [58,60] evident that when the number of hidden neurons is increased, the MSE and RMSE decrease for both training and testing data. However, after a certain point, the MSE and RMSE decrease for training but increase for testing. This study also indicates that after a certain point, the error in testing data will continue to rise without much change in the training data. To ensure that the training, testing, and validation sets are evenly distributed across different classes and to avoid any potential issues that may arise from having similar or sequential data in the sets, the optimal number of neurons for the ANN model was determined by randomizing the pooled dataset. This resulted in a best-hidden layer size of 7. In the context of a Backpropagation Neural Network, randomization helps to prevent the algorithm from quickly converging to local minima by introducing oscillations.

To optimize the ANN, the Min-Max transformation is applied to input and output values to speed up the training process, avoid saturation, and reduce the chances of getting stuck in local optima. This transformation shifts the data into the range [–1, 1] through the following equation:

$$X_{i,scaled} = \frac{X_i - X_{min}}{X_{max} - X_{min}}$$

Where X_i is the original input value. After the ANN simulation, the transformed values are converted back to their original values using various transfer functions. Johnstone and Sulungu [61] discuss the most commonly used transfer functions, such as linear, hyperbolic tangent sigmoid, and logistic sigmoid, which are suitable for problems with nonlinearity. These transfer functions enable the ANN to accurately convert the transformed values back to their original form. In this study, a feed-forward backpropagation neural network with 7 hidden neurons and a Hyperbolic tangent sigmoid transfer function was implemented using the MATLAB ANN toolbox and the Levenberg Marquardt training algorithm for deep learning of summer Tmax over Taiwan. This choice of the transfer function is effective for temperature prediction, as it is a nonlinear, differentiable, and monotonic function that yields better training performance for multilayer neural networks. In this study, a basic ANN is created using the components outlined in Table 1.

Table 1. The following are considered to develop a simple ANN model to improve the GEFSv12 prediction skill in depicting Summer daily Tmax and Associate Tmax extremes over Taiwan.

No. Hidden layers :	1
No. of nodes/neurons in the hidden layer	7
Neural Network used	Feedforward network
Neural Network Processing Functions	Map matrix row minimum and maximum values to [–1 1]
Data divided function	70% data for training and 30% data for validation
Learning rate	0.001

Max number of iterations/epochs used	1000
Error tolerance for stopping criterion	1e-14
Training function used	Supervised weight/bias training function with Sequential order weight/bias training (trains)
Neural Network Performance Functions used	Mean squared error performance function

For summer T$_{max}$ forecasts over Taiwan, the ANN calibration method is applied to each forecast lead time using a 31-day moving window with 620 sample data (31 days * 20 years). The forecast day is the center of the window. The LOOCV procedure has been implemented to calibrate the outputs of the GEFSv12 using ANN. The resulting calibrated outputs are referred to as ANN-GEFSv12.

2.2.3. Hybrid Post-processing:

The QQ technique, described in Section 2.2.1, was used to improve summer daily Tmax and associated Tmax extremes in Taiwan by applying it to the ANN-GEFSv12 output. The LOOCV procedure was used to evaluate the performance of the hybrid statistical post-processing method. A visual representation of the methodology used in this study is presented in Figure 1. The predictive accuracy of the different calibration methods for deterministic and ensemble probabilistic forecasts of Tmax and associated extremes was compared using standard skill metrics.
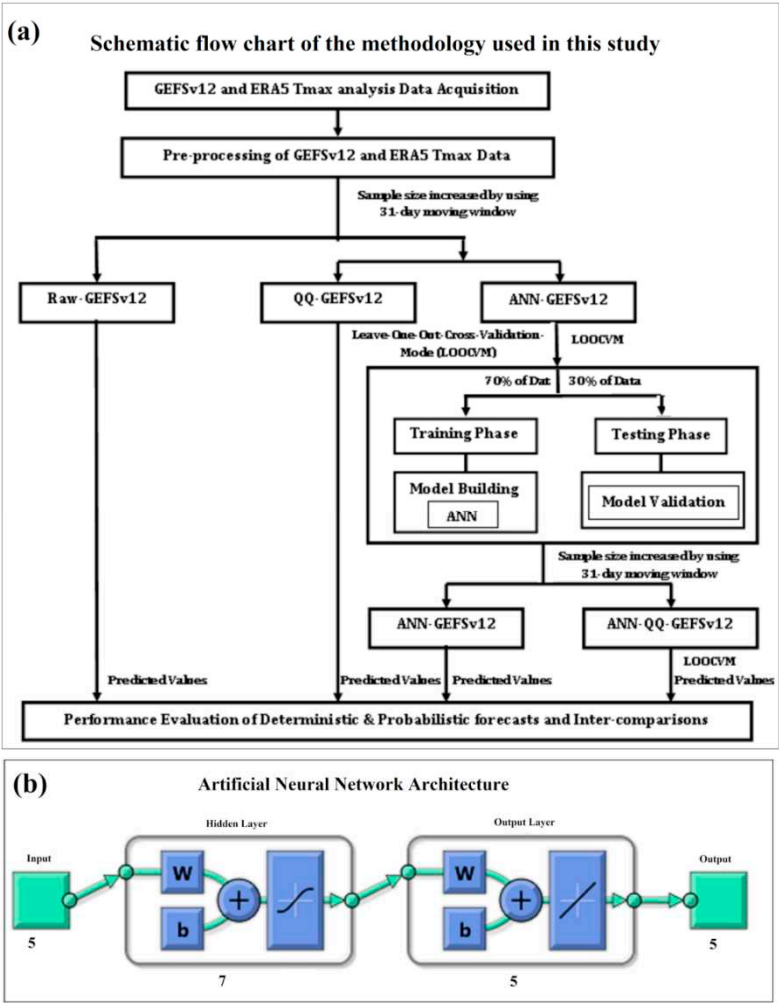
7

Figure 1. (a) Schematic flow chart of methodology and **(b)** Artificial Neural Network Architecture used in this study.

2.3. Analysis Procedure:

The accuracy of Raw, QQ, ANN, and Hybrid methods in predicting summer daily Tmax over Taiwan for day-1 to 16 forecast lead times during the reforecast period (2000-2019) was evaluated against ERA5 using skill metrics such as mean bias (MB), Root Mean Square Error (RMSE), correlation coefficient (CC), and index of agreement (IOA). The probability distributions of the Raw and all three calibration methods were compared with ERA5 by pooling all grid points and 5 ensemble members for each forecast lead time separately. The spatial distribution of summer (JJAS) Tmax extremes over Taiwan was analyzed, taking into account the average frequency of Tmax extremes from all five individual members for each forecast lead time separately. The performance of Raw and all three calibration methods in predicting summer Tmax extremes against ERA5 was evaluated using a contingency table and associated statistical categorical skill scores, such as Accuracy (ACC), Frequency Bias (BIAS), Probability of Detection (POD), False Alarm Rate (FAR), Success Ratio (SR), Threat Score (TS), and Equitable Threat Score (ETS). A performance diagram has been created to illustrate the statistical categorical skill scores of Raw and all calibration methods in depicting summer daily Tmax extremes. This diagram measures the geometric relationship between Frequency Bias, SR, FAR, POD, and TS [62].

Probabilistic forecasts are essential for providing more accurate and reliable weather and climate predictions, as they are better able to capture the inherent uncertainty of extreme events. To evaluate their accuracy, metrics such as Reliability, Resolution, Brier score (BS), Brier skill score (BSS), and Receiver operating characteristic (ROC) curve are used. These metrics are invaluable for climate risk management in various sectors. The Brier Score is a measure of the accuracy of probabilistic forecasts in binary situations, ranging from 0 to 1, with 0 being the perfect score. The Brier Skill Score (BSS) is used to compare the accuracy of a probabilistic forecast to a reference/climatological forecast. A BSS of 1 indicates an accurate forecast, while a BSS of 0 or lower suggests that the forecast is not as reliable as the reference [63,64]. The reliability and resolution of an ensemble probabilistic forecast of a particular category are two distinct characteristics [65]. The reliability of an ensemble probabilistic forecast is the accuracy of the predicted class/interval of outcomes compared to the actual distribution of observations. A perfectly calibrated forecast has a reliability of 0, while a scale from 0 to 1 measure the reliability of a forecast and 1 representing the worst reliability. The resolution of a forecast is a measure of its accuracy in predicting the frequency of an event. A resolution of 0 indicates that the forecast is either always the same or completely random, while a resolution equal to the uncertainty means that all uncertainty has been accounted for. The Receiver Operating Characteristic (ROC) curve plots the False Alarm Rate (FAR) on the x-axis against the Probability of Detection (POD) on the y-axis. A forecast with skill will have a curve above the diagonal line, while a forecast below the line is worse than a climatological or reference forecast. An accurate forecast will be close to the ideal upper left corner [66].

3. Results

This study applied various calibration methods to NOAA NCEPGEFSv12 reforecasts to improve the predictability of summer daily Tmax and associated Tmax extremes over Taiwan. The performance of these methods was evaluated using standard skill metrics for deterministic and ensemble probabilistic forecasts. The results are discussed in the following subsections.

3.1. Prediction skill of Raw, QQ, ANN, and Hybrid post-processing methods for summer daily Tmax over Taiwan

The performance of raw and three calibration methods (ERA5, QQ, ANN, and Hybrid) for predicting summer (JJAS) daily Tmax over Taiwan for 2000-2019 was evaluated by analyzing the spatial patterns of the climatological mean at forecast lead times of Day-1, 5, 10, and 15 (Figure 2). Both Raw-GEFSv12 and ERA5 show similar spatial patterns of summer daily Tmax over Taiwan for all

forecast lead times However, GEFSv12 has a warm bias in most parts of the country. The daily T_{max} from ERA5 is lower in the east and increases towards the west, which is also seen in GEFSv12. The highest summer T_{max} is seen in the southernmost region of Taiwan. The GEFSv12 forecasts for all lead times reflect this. All calibration methods notably reduced the warm bias in most parts of Taiwan, resulting in a T_{max} climatological mean similar to ERA5 for all forecast lead times.
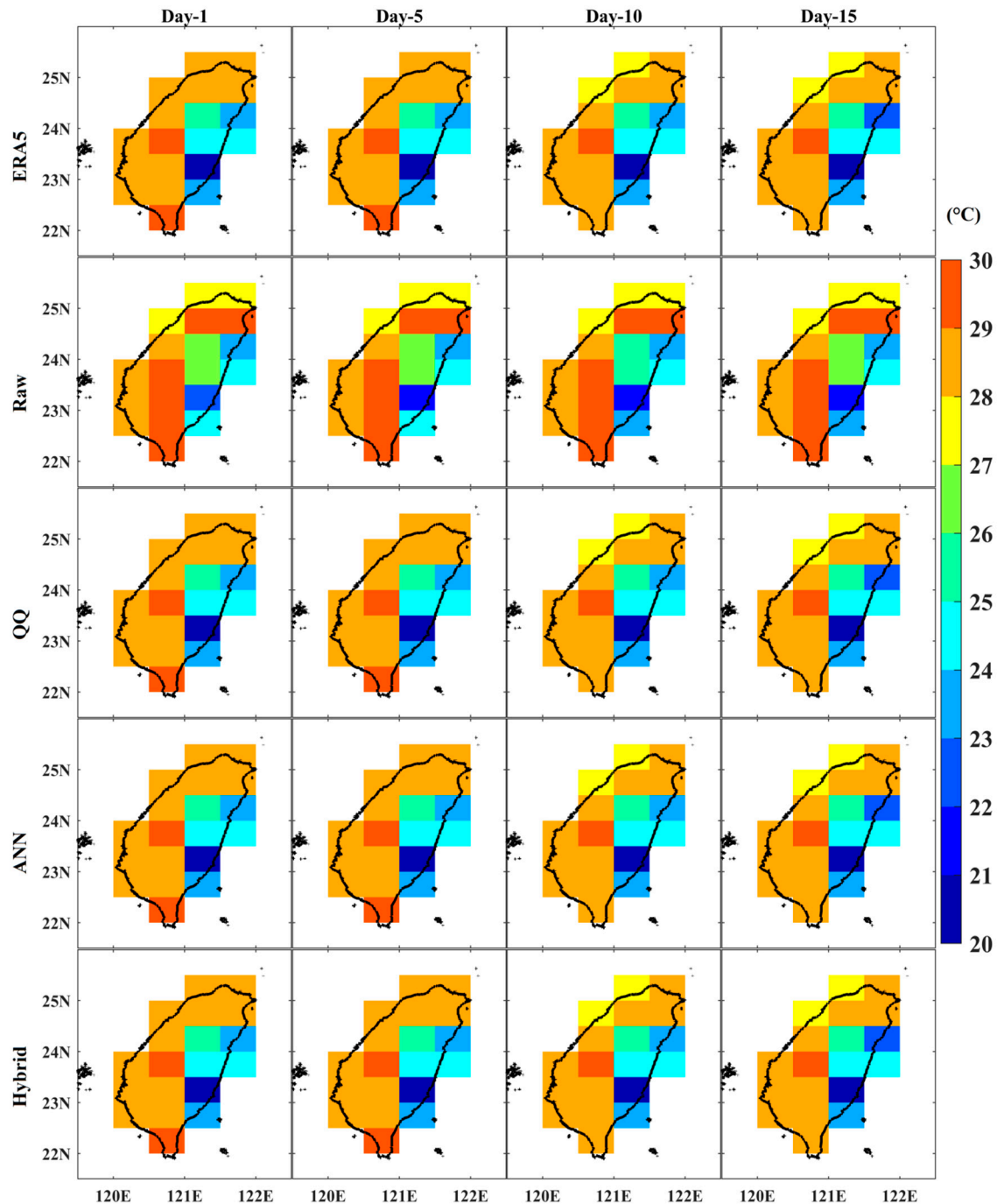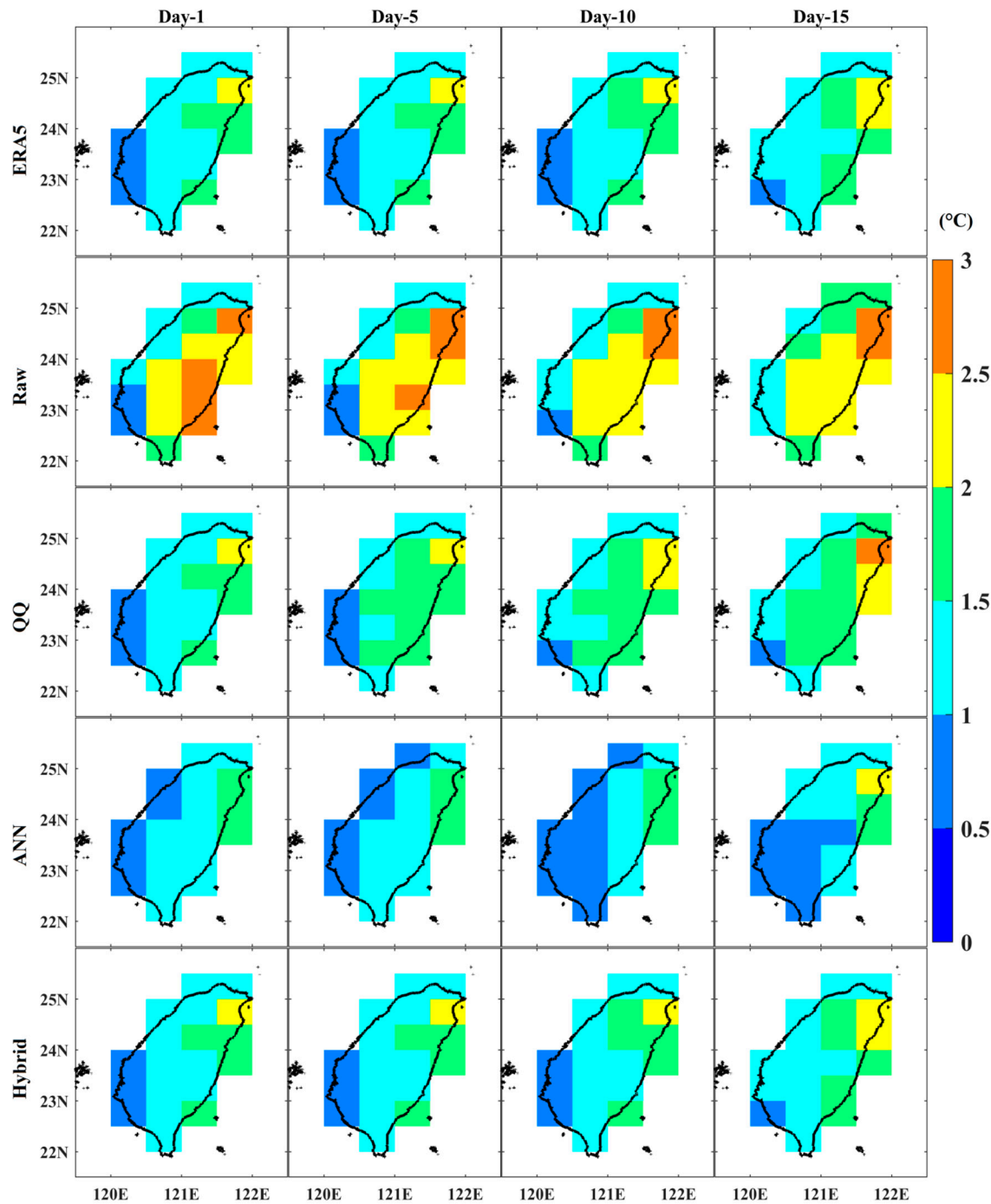

Figure 2. Climatological mean of summer (JJAS) surface air maximum Temperature (T_{max}) over Taiwan from ERA5, Raw, QQ, ANN, and Hybrid methods with Day-1, 5, 10, and 15 forecast lead times for the period 2000-2019.

The spatial patterns of IAV of summer T_{max} over Taiwan from GEFSv12 and ERA5 are similar for all forecast lead times (Figure 3). GEFSv12 tends to overestimate the IAV of summer T_{max} in most parts of the country for all forecast lead times. The IAV of T_{max} is higher in the northeastern part of the country, which is accurately reflected in the GEFSv12 forecasts for all lead times. All three

calibration methods successfully reduced the overestimation of T$_{max}$ IAV over Taiwan. The spatial patterns of the IAV of T$_{max}$ were found to be similar to those of ERA5 for all forecast lead times. The ANN method slightly underestimated the IAV of T$_{max}$ in most parts of the country, while the QQ and Hybrid methods accurately captured the magnitude of the IAV of T$_{max}$ over Taiwan for all forecast lead times. The Hybrid method of capturing T$_{max}$ IAV in Taiwan is more effective than the QQ method, especially for longer lead time forecasts (Figure 3).

Figure 3. Interannual variability of summer (JJAS) surface air maximum Temperature (T$_{max}$) over Taiwan from ERA5, Raw, QQ, ANN, and Hybrid methods with Day-1, 5, 10, and 15 forecast lead times for the period 2000-2019.

The QQ method has the advantage of adjusting the T_{max} probability distribution to the observed data, particularly in the extreme tails, to account for IAV. The spatial patterns have been improved, however, the temporal patterns remain the same. Deep learning combined with the QQ method has been found to be effective in capturing temporal patterns, IAV, and climatological patterns. The Hybrid method has been seen to be more successful than the QQ and ANN methods.

The Raw-GEFSv12 model showed a high RMSE in predicting summer daily T_{max} in the eastern parts of Taiwan for all forecast lead times (Figure 4). The RMSE patterns were similar to the IAV patterns, with higher values in high IAV regions. The RMSE increased with lead time. All three calibration methods effectively reduced the RMSE in most parts of Taiwan for all forecast lead times. The RMSE of the QQ method is higher for longer lead times, while the ANN and Hybrid methods show significant improvements. The comparison between the methods reveals that the RMSE of ANN and Hybrid methods is lower than that of the QQ method for all forecast lead times, particularly in the eastern parts of the country (Figure 4).
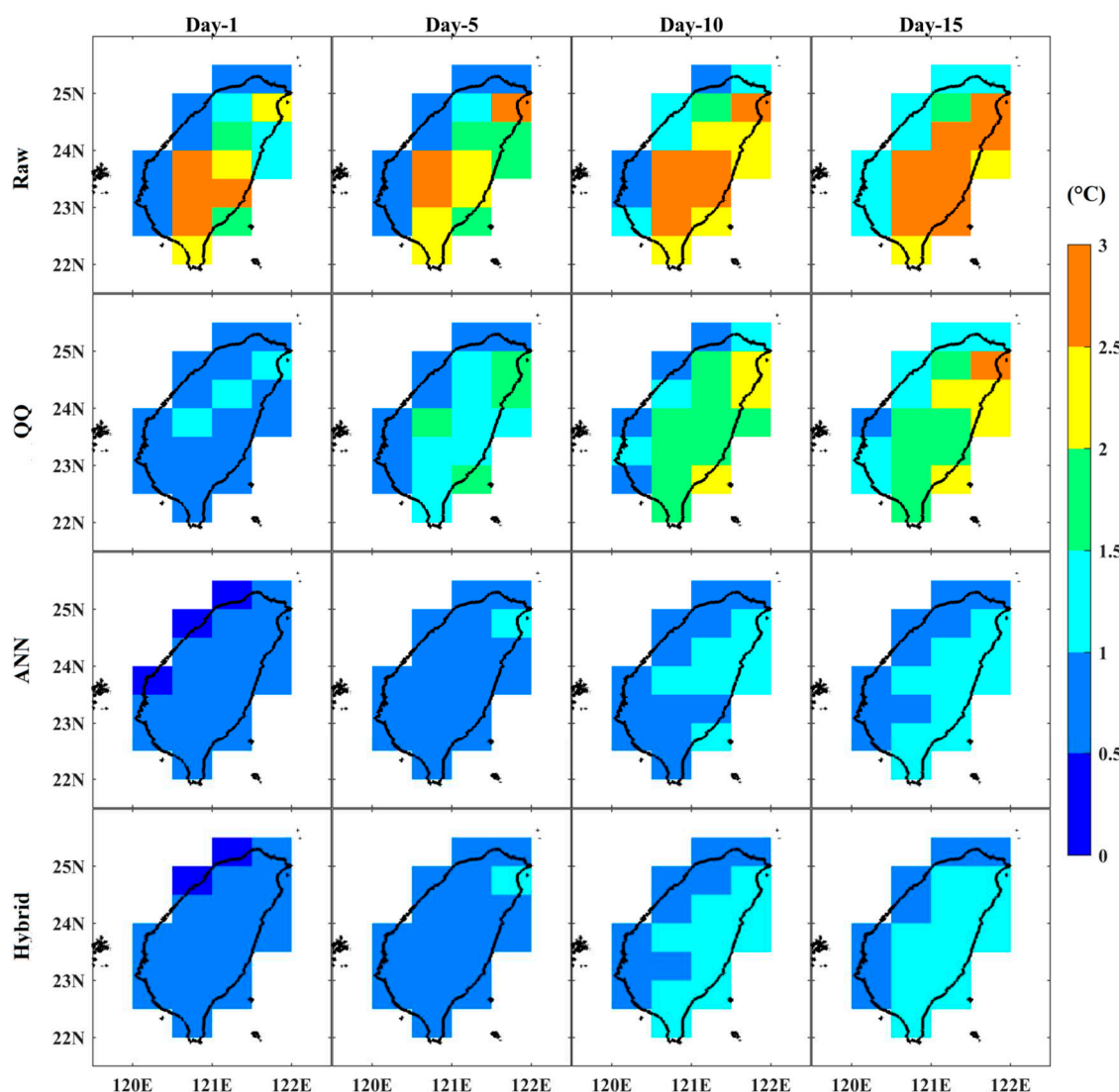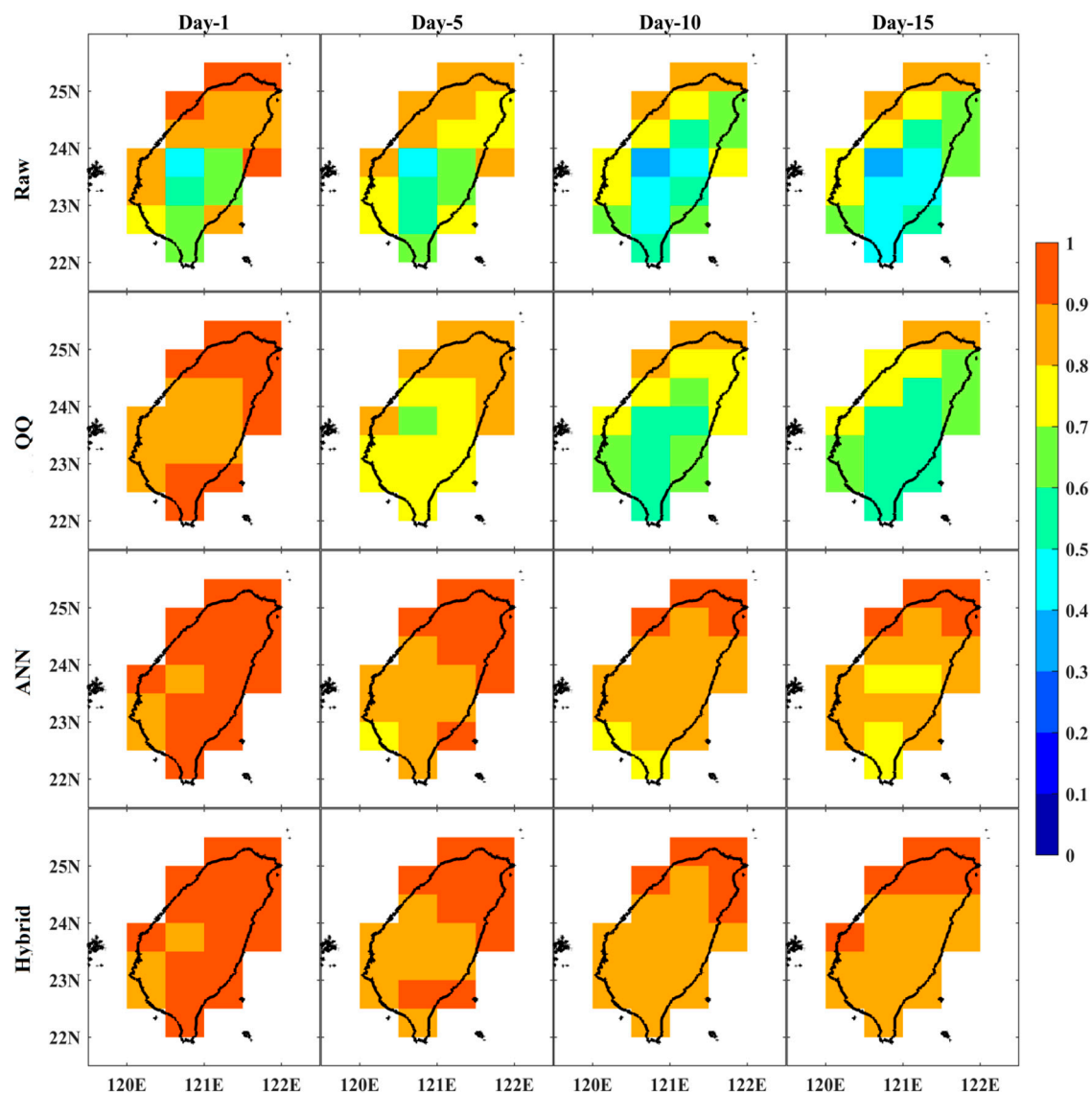

Figure 4. RMSE (°C)of Raw, QQ, ANN, and Hybrid methods with Day-1, 5, 10, and 15 forecast lead times against ERA5 in depicting summer (JJAS) surface air maximum Temperature (T_{max}) over Taiwan for the period 2000-2019.

GEFSv12 shows a high Index of Agreement (IOA) (> 0.8) for predicting summer daily T_{max} in northwestern Taiwan, decreasing to > 0.5 in the southeast (Figure 5). However, the IOA is lower in the central part of the country for all forecast lead times. The IOA of GEFSv12 for summer daily T_{max}

generally decreases with increasing forecast lead time in most areas. However, the application of calibration methods has significantly improved the IOA of predicting T_{max} over Taiwan for all forecast lead times. The ANN method has an IOA range of 0.7 to 1, which is higher than the QQ range of 0.5 to 1. The accuracy of the forecasts for T_{max} in all parts of Taiwan produced by ANN is significantly higher for longer lead times. On the other hand, the IOA from QQ decreases with increasing lead time, mainly due to larger errors in the forecasts. The Hybrid method, however, has a higher IOA value (0.8-1) than the other two methods, making it the most reliable for predicting summer daily T_{max} over Taiwan for all forecast lead times. Hybrid methods of predicting T_{max} demonstrate more reliable results across the majority of the country compared to ANN and QQ for all forecast lead times (Figure 5).

Figure 5. Index of Agreement of Raw, QQ, ANN, and Hybrid methods with Day-1, 5, 10, and 15 forecast lead times against ERA5 in depicting summer (JJAS) surface air maximum Temperature (T_{max}) over Taiwan for the period 2000-2019.

The performance of the Raw and all three calibration methods in predicting T_{max} over Taiwan for the reforecast period was evaluated using RMSE, Mean Bias, Correlation Coefficient, and Index of Agreement (Figure 6). Results showed that the RMSE increased with increasing forecast lead time.

The highest RMSE was observed for the Raw, ranging from 1.5 to 2.5°C. However, the application of calibration methods such as QQ (0.8-1.2°C), ANN (0.6-1°C), and Hybrid (0.6-1°C) significantly reduced the RMSE for all forecast lead times (Figure 6a). The comparison of the methods reveals that ANN and Hybrid have similar RMSE values, which are much lower than QQ for all forecast lead times (Figure 6b). The warm bias of 0.6-1°C over Taiwan during the summer season was successfully reduced to nearly 0°C by all calibration methods. GEFSv12 shows a strong correlation with summer daily T_{max} over Taiwan for Day-1 forecasts (r>0.8), decreasing with increasing lead time (r=0.4) (Figure 6c). No improvement was seen in the correlation coefficient when using the QQ method compared to the Raw products. However, the ANN and Hybrid calibration methods both showed a significant improvement in the correlation coefficient (r>0.79) for all forecast lead times. The Hybrid method yields the same correlation coefficient values as the ANN for all forecast lead times. However, for longer lead times, both the ANN and Hybrid methods show a significant improvement in the correlation coefficient (Figure 6c). The IOA of GEFSv12 in predicting T_{max} over Taiwan is highest for shorter lead times (0.8) and decreases to 0.6 as the forecast lead time increases (Figure 6d). All the calibration methods improve the IOA for all forecast lead times. The Hybrid method shows the highest IOA (0.92) compared to the ANN (0.88) and QQ (0.9). The Hybrid method yielded higher IOA values than the ANN for all forecast lead times (Figure 6d). The QQ calibration method also had lower IOA values than the Hybrid method for all forecast lead times (Figure 6d). This improvement in accuracy is especially beneficial for longer lead time forecasts, which can be immensely helpful for climate management in various sectors at the regional level, such as Taiwan.
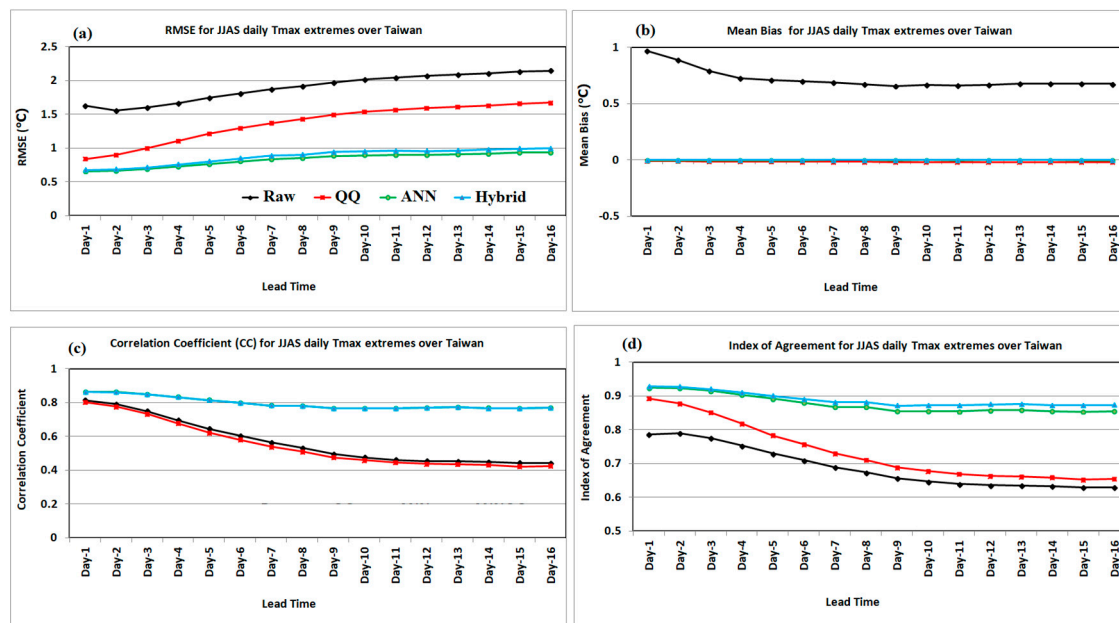

Figure 6. (a) Root mean squared error in°C, (b) Mean Bias in °C, (c) Correlation Coefficient, and (d) Index of Agreement of Raw, QQ, ANN, and Hybrid methods against ERA5 in depicting Summer daily T_{max} over Taiwan for Day-1 to 16 lead time forecasts for the period 2000-2019.

The probability distribution (PDF) of summer daily T_{max} over Taiwan was calculated from all 5 ensemble members and all grid points of Taiwan daily T_{max} values pooled for ERA5, Raw and each calibration method for the study period and selected lead time forecasts (Day-1, 5, 10, and 15). The results are shown in Figure 7. The PDF of the summer daily T_{max} from Raw is right-skewed compared to the ERA5 for all forecast lead times. This indicates that the number of extreme days with higher T_{max} is higher in Raw data than in ERA5 data. The calibration methods were well-adjusted for the probability distribution of summer daily T_{max} over Taiwan to ERA5 for all the forecast lead times. The QQ method was found to be more effective than the ANN. The Hybrid method was found to be the

most effective in adjusting the PDF of summer daily T_{max} over Taiwan to ERA5. The Hybrid method outperforms the QQ and ANN in adjusting the summer daily T_{max} PDF to ERA5.
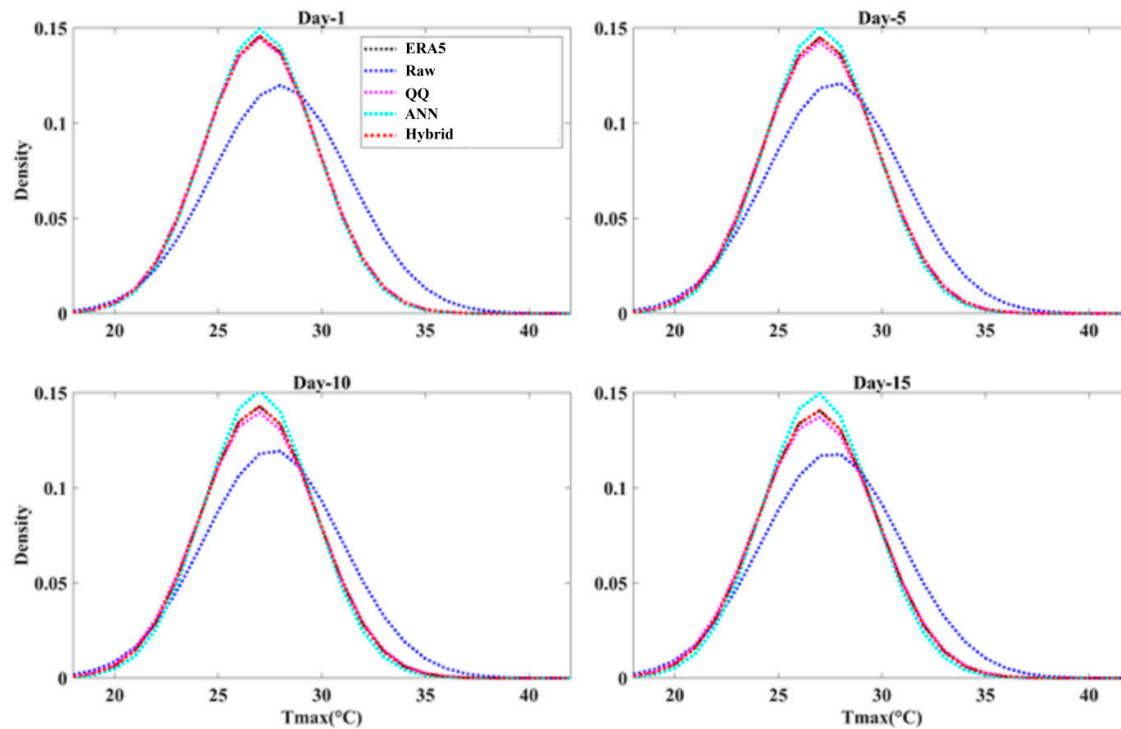

Figure 7. The PDF of summer (JJAS) daily T_{max} over Taiwan from ERA5 (black dotted lines), Raw (Blue dotted lines), QQ (Magenta dotted lines), ANN (Cyan dotted lines), and Hybrid methods (Red dotted lines) for Day-1,5,10, and 15 forecast lead times for the reforecast period 2000-2019.

3.2. Statistical Categorical Skill Scores for Summer Daily Tmax Extremes over Taiwan from Raw, QQ, ANN, and Hybrid Methods

Statistical skill scores (e.g. POD, FAR, ACC, SR, TS, ETS) were computed for the 2000-2019 reforecast period for Taiwan's summer daily T_{max} extreme days (T_{max} > 90th percentile of annual T_{max}) from day-1 to 16. The ETS of GEFSv12 for summer daily T_{max} extremes is higher in coastal areas than in interior regions of Taiwan (Figure 8). The ETS values decrease with increasing forecast lead time. All calibration methods tested showed an improvement in the ETS score for summer daily T_{max} extremes over Taiwan for all forecast lead times. Raw and all three calibration methods for summer daily T_{max} over Taiwan indicate a decrease in ETS score with increasing forecast lead times. However, the ANN method yields a higher ETS score than the QQ calibration method. The Hybrid method yields the highest ETS score than ANN and QQ for all forecast lead times (Figure 8).
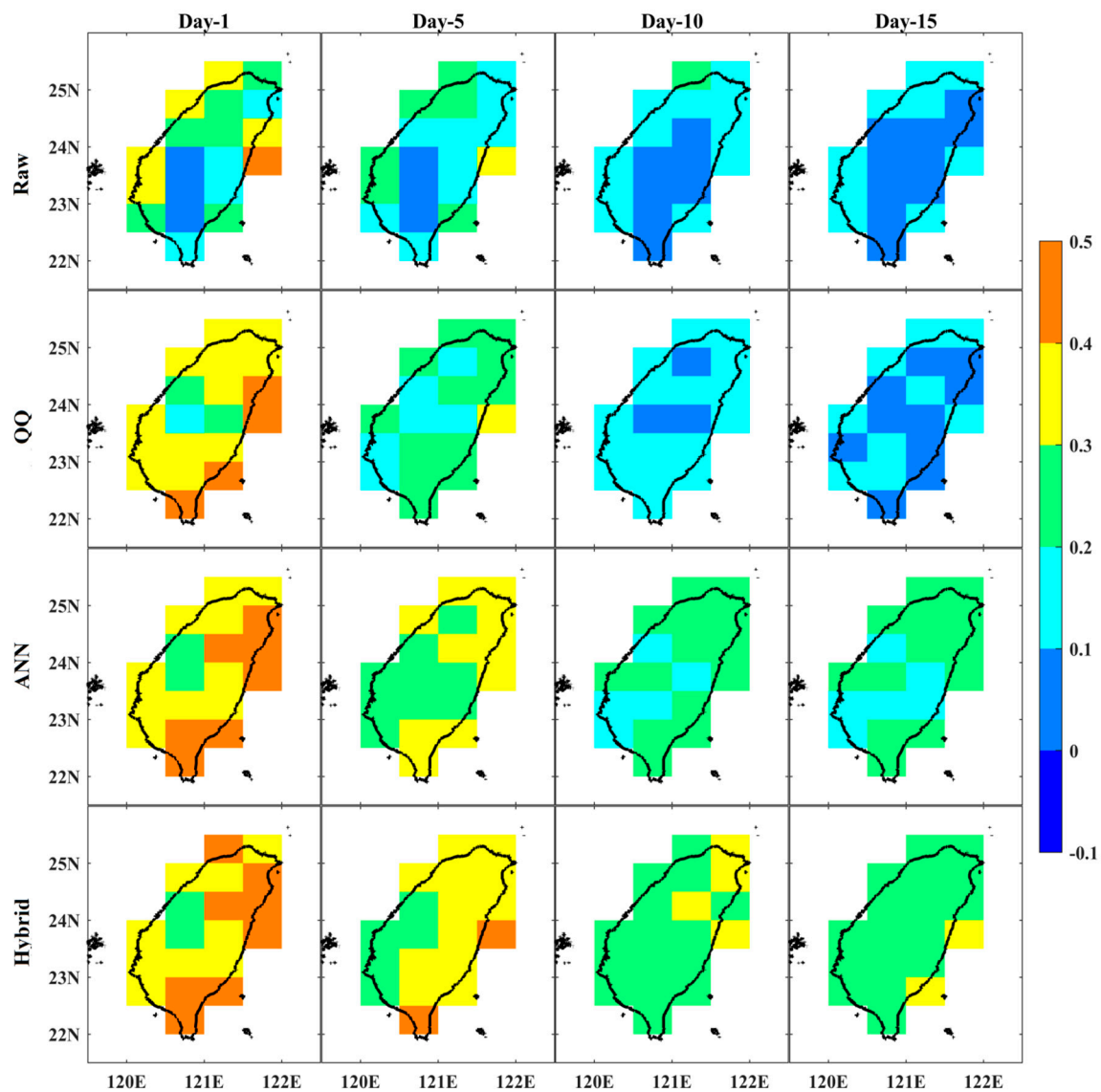
Figure 8. The Equitable Threat Score (ETS) of Raw, QQ, ANN, and Hybrid methods in depicting JJAS daily Tmax extremes over Taiwan against ERA5 with Day-1, 5, 10, and 15 forecast lead times for the period 2000-2019.

The ETS scores for the Week-1, Week-2, and Week-1 to 2 scales were further analyzed. Results showed that the Hybrid method had the highest ETS score for all forecast lead times. The ETS score for predicting summer daily Tmax extremes over Taiwan from GEFSv12 is higher for Week-1 than Week-2, as seen in Figure 9. The ETS score from GEFSv12 for the two-week period (Week-1 to Week-2) is higher than the ETS scores of Week-1 and Week-2 for predicting summer daily Tmax extremes in Taiwan. All three calibration methods improve the ETS score for summer daily Tmax extremes for Week-1, Week-2, and Week-1 to 2. The ETS score of summer daily Tmax from all three methods is higher for Week-1 than Week-2 and Week-1 to Week-2. The comparative analysis shows that the ETS from ANN in most parts of Taiwan for summer daily Tmax extremes for Week-1, 2, and 1 to 2 is relatively higher than the QQ calibration method. The Hybrid method for summer daily Tmax extremes for Week-1, 2, and 1 to 2 yielded notably higher ETS scores in most parts of Taiwan than the ANN and QQ calibration methods.
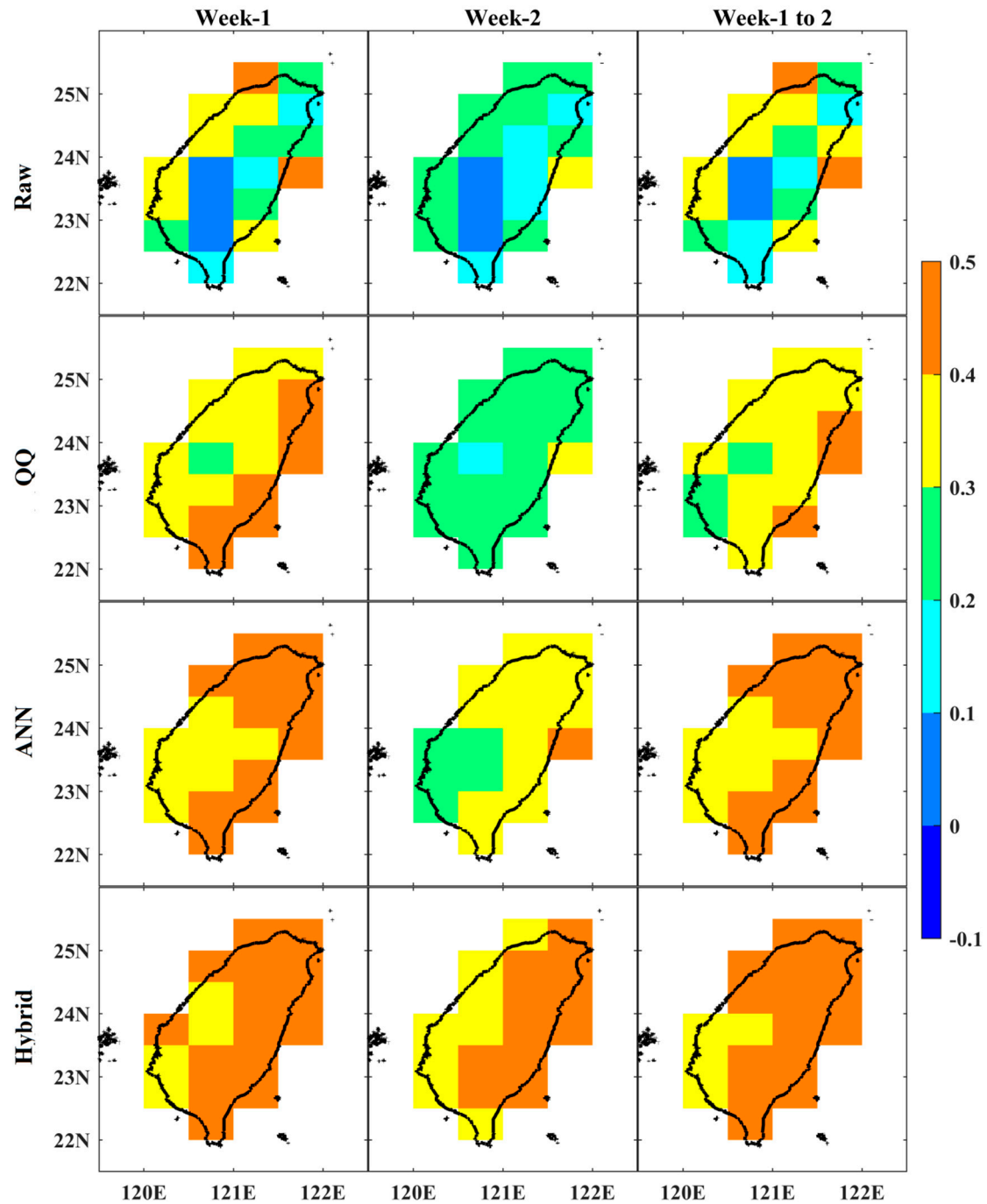
Figure 9. The Equitable Threat Score (ETS) of Raw, QQ, ANN, and Hybrid methods in depicting JJAS daily T$_{max}$ extremes over Taiwan against ERA5 for Week-1, 2, 1 to 2 forecasts for the period 2000-2019.

A performance diagram is a graphical representation of multiple skill scores, such as POD, Frequency Bias, TS, and SR (1-FAR), which can be used to compare and analyze performance [62]. Figure 10a shows that the GEFSv12 model overestimates summer daily T$_{max}$ extreme days over Taiwan for all forecast lead times, with a Frequency Bias of more than 1.5 and a POD ranging from 0.6 to 0.8. The SR and TS scores of GEFSv12 decreases with increasing forecast lead time. However, the three calibration methods have been found to effectively reduce the overestimation of daily T$_{max}$ extremes over Taiwan for all forecast lead times. For summer daily T$_{max}$ extremes over Taiwan, the POD has decreased for all forecast lead times when using all three calibration methods. However, the QQ method showed higher POD values than ANN for longer lead time forecasts.
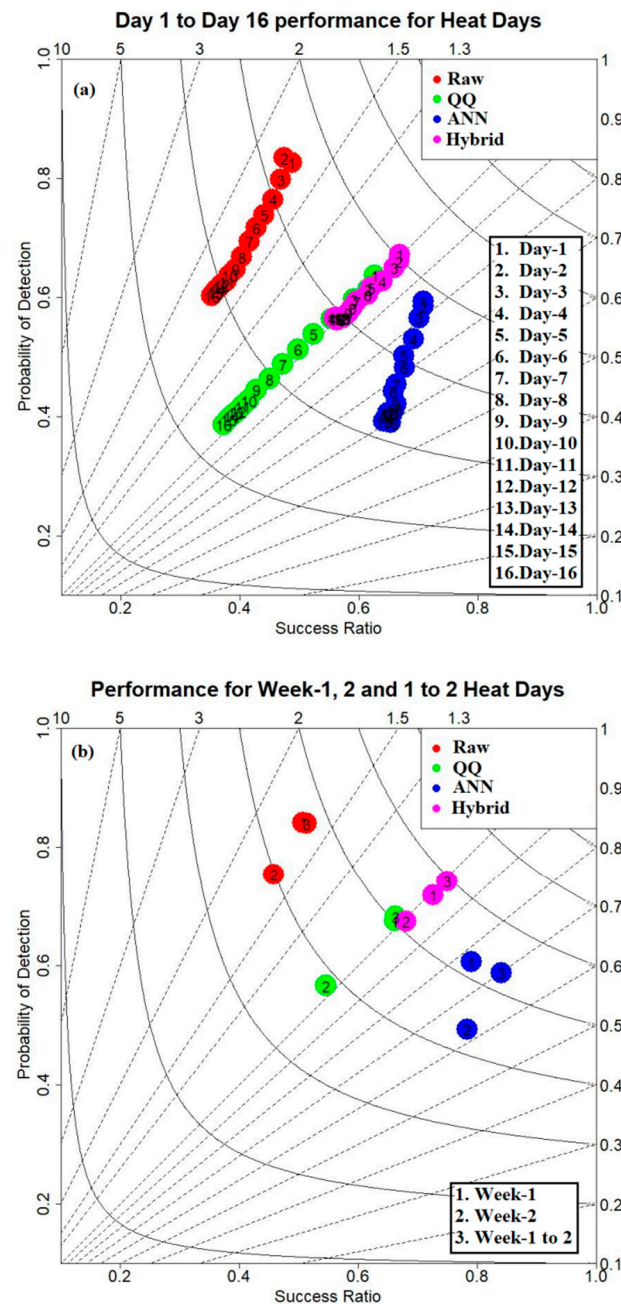
Figure 10. Performance diagram illustrating the SR, POD, Frequency Bias, and TS statistical categorical skill scores of Raw, QQ, ANN, and Hybrid methods against ERA5 for JJAS daily Tma T_{max} extremes over Taiwan on (a) a daily scale with Day-1 to 16, and (b) Weekly Scale for Week-1, 2, 1 to 2 for the period 2000-2019 is presented, with solid and dashed lines representing TS and Frequency Bias scores, respectively.

The ANN model yields higher SR and TS values than the QQ method for all forecast lead times. Both the QQ and Hybrid calibration methods are able to accurately reproduce the number of summer T_{max} extreme days observed in ERA5. However, the Hybrid method outperforms the other two methods in terms of POD, SR, and TS skill scores. This suggests that the Hybrid method could be beneficial for extended-range time-scale predictions.

The comparison of GEFSv12 with three calibration methods for Week-1, 2, and 1 to 2 revealed a substantial overestimation of summer T_{max} extreme days (Figure 10b). All three calibration methods were successful in reducing the overestimation. However, the Hybrid method showed the highest statistical categorical skill scores. The skill scores from Raw, QQ, ANN, and Hybrid calibration

methods were generally higher for Week-1 and Week-1 to 2 than for Week-2 (Figure 10b). This suggests that the GEFSv12 summer T_{max} extreme day data is not reliable without calibration. The Hybrid method was found to be the most effective in improving the skill scores for all forecast scales. This makes it a valuable tool for climate risk management in the region.

3.3. Probabilistic Prediction Skill Scores of Raw, QQ, ANN, and Hybrid methods for Summer daily Tmax Extremes

The uncertainty of summer T_{max} extremes over Taiwan can be evaluated using metrics such as Resolution, Reliability, Brier Score, Brier Skill Score, and ROC curves to assess the ensemble probabilistic forecast. The GEFSv12 probabilistic forecast of summer T_{max} extreme days over Taiwan has a good reliability (< 0.15) for all forecast lead times, as shown in Figure 11a. This was further improved by the application of three calibration methods (< 0.05). The reliability of the forecast decreases with increasing lead time from Raw and all three calibration methods. However, the ANN and Hybrid methods showed the highest reliability, particularly for longer lead time forecasts. The resolution of the GEFSv12 model for probabilistic forecasts of summer T_{max} extreme days over Taiwan decreases with increasing forecast lead time, with higher resolution for shorter lead times (Figure 11b). All three calibration techniques significantly improved the resolution of the ensemble probabilistic forecast of summer T_{max} extreme days over Taiwan for all forecast lead times. ANN and Hybrid methods showed the highest resolution. Especially for longer lead times, the resolution of ANN and Hybrid methods was significantly higher. The hybrid calibration method has a relatively better resolution than the ANN for all forecast lead times (Figure 11b). The Brier Score (BS) is a metric used to measure the accuracy of binary predictions, where the result is either yes or no. The ideal score is 0. According to Figure 11c, the confidence of GEFSv12's ensemble probabilistic forecasts of summer T_{max} extreme days over Taiwan is low (BS > 0.25) for all forecast lead times. However, the calibration methods used were found to be highly effective in improving the accuracy (BS < 0.2) of these forecasts. Specifically, the ANN and Hybrid calibration methods showed higher accuracy than the QQ method. The Hybrid method of ensemble probabilistic forecasting of summer T_{max} extreme days over Taiwan produces results similar to those of the ANN for all forecast lead times (Figure 11c). The GEFSv12 ensemble probabilistic forecasting of summer T_{max} extreme days over Taiwan with a BSS of less than -0.4 was not as accurate as the climatological/random forecast for all forecast lead times. This was evident from the results shown in Figure 11d. However, the use of calibration methods such as QQ, ANN, and Hybrid methods improved the BSS remarkably for all forecast lead times. The QQ method was found to be the most accurate for up to one week lead time than the reference/climatological/random forecast. After the first week, the ensemble probabilistic forecasting of summer T_{max} extreme days over Taiwan from QQ was not as accurate as expected as the random forecast. However, the ANN and Hybrid methods outperformed both the random forecast and QQ for all forecast lead times (Figure 11d). The Hybrid method is more effective than ANN for predicting extreme summer T_{max} days over Taiwan for all forecast lead times.
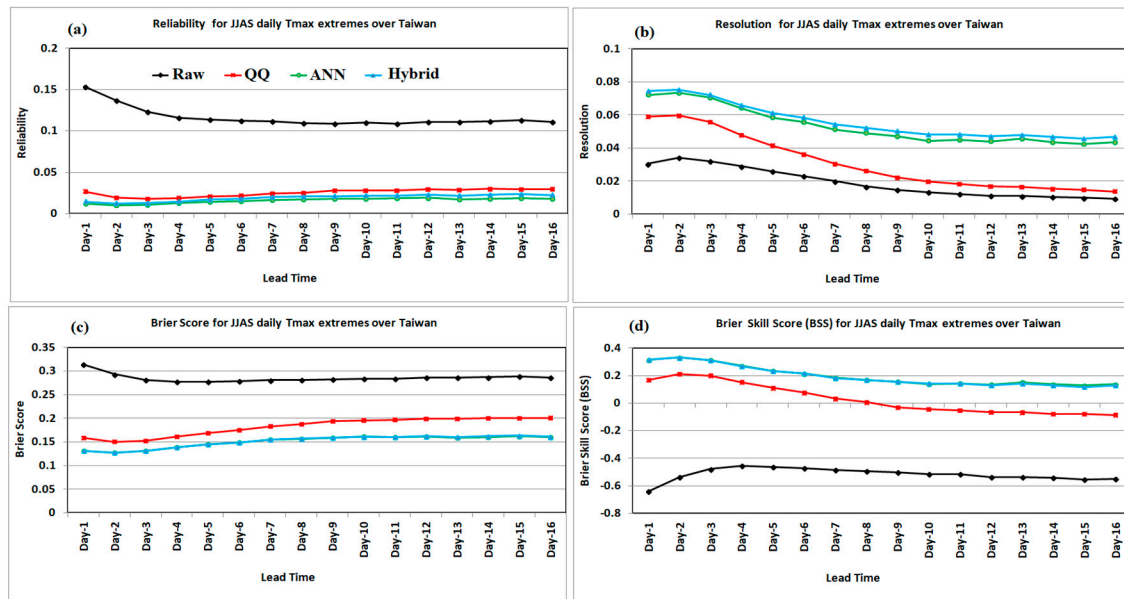
Figure 11. (a) Reliability, (b) Resolution, (c) Brier score and (d) Brier skill score of Raw, QQ, ANN, and Hybrid methods against ERA5 for summer daily T_{max} extremes Ensemble probabilistic forecast over Taiwan with Day-1 to 16 forecast lead times for the period 2000-2019.

As a final diagnostic, we use the ROC curve to assess a model's ability to distinguish between events and non-events (Wilks, 2011). The ROC curve evaluates the forecast if a summer T_{max} extreme day had occurred. It plots the true positive rate (correctly predicted T_{max} extreme day) against the false positive rate (incorrectly predicted T_{max} extreme day). We calculate the true positive rate and false positive rate for cumulative probabilities ranging from 0% to 100%, in increments of 10%. A skillful forecasting model should have a higher true positive rate than a false positive rate, resulting in a ROC curve that curves towards the top-left corner of the plot. Conversely, a forecast system with no skill would be a straight line along the diagonal, indicating that the forecast is no better than a random guess. The AUC (Area Under the Curve) is a useful scalar measure for summarizing the performance of a model, with a score of 1 indicating the highest level of skill and a score of 0 indicating the lowest level of skill. The ROC curves for Raw, QQ, ANN, and Hybrid calibration methods for ensemble probabilistic forecasting of summer T_{max} extreme days over Taiwan are all above the diagonal line for all forecast lead times, as shown in Figure 12.

Raw and all three calibration methods for ensemble probabilistic forecasting of summer T_{max} extreme days over Taiwan have a satisfactory AUC skill score (> 0.65) for all forecast lead times. However, it has been observed that the AUC skill decreases with increasing forecast lead times. The Hybrid calibration method yielded the highest AUC skill score (0.79-0.85), followed by ANN (0.75-0.83), QQ (0.68-0.81), and Raw (0.65-0.74). The performance analysis of three calibration methods revealed that they significantly improved the accuracy of GEFSv12 in forecasting extreme summer T_{max} days in Taiwan. The Hybrid calibration method for ensemble probabilistic forecasting of summer T_{max} extreme days on an extended range time scale over Taiwan has been shown to be more effective than the QQ and ANN techniques.
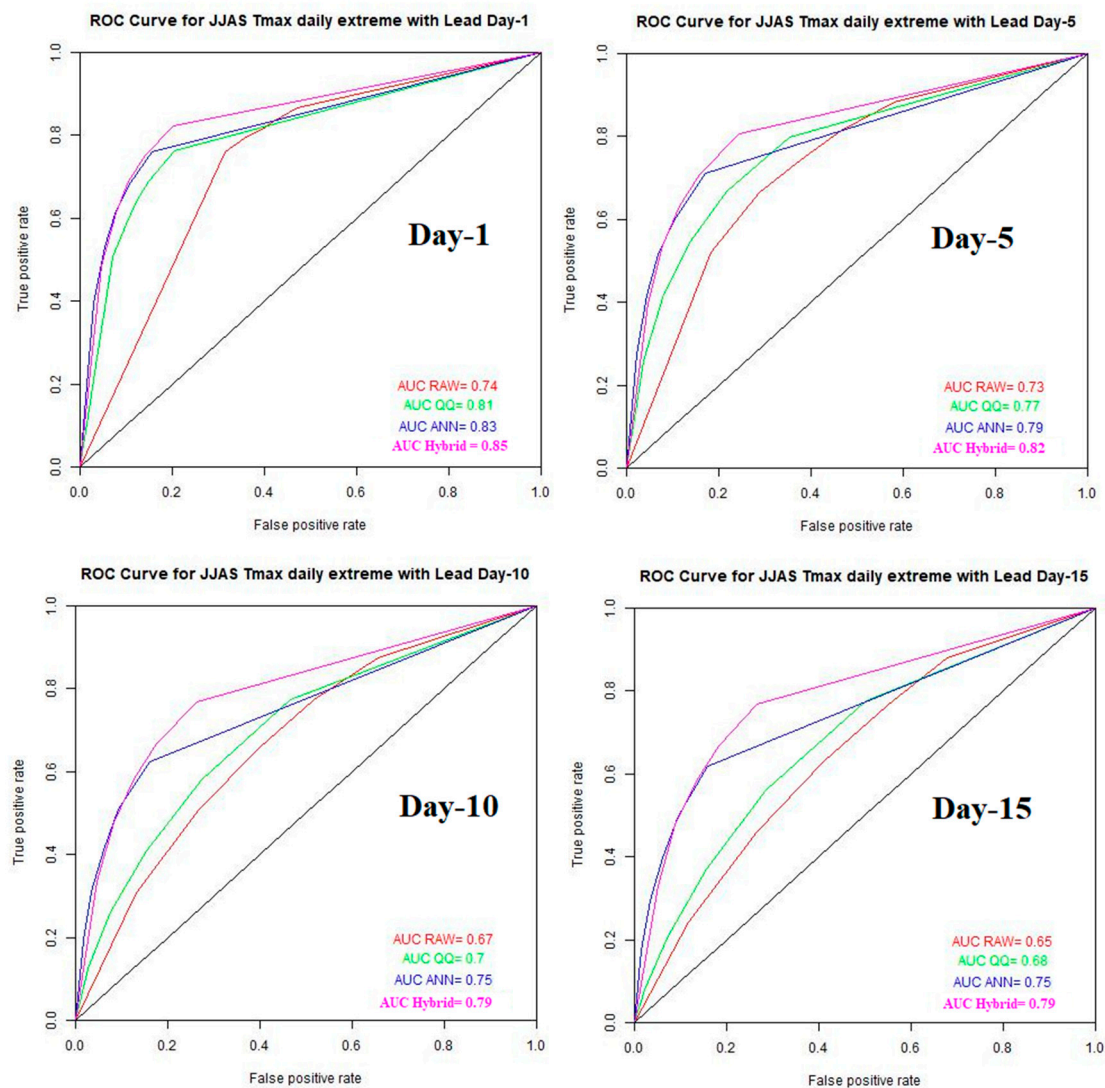
Figure 12. Receiver operating characteristic (ROC) curve and Area under the ROC Curve of Raw, QQ, ANN, and Hybrid methods against ERA5 for summer extreme daily T$_{max}$ Ensemble probabilistic forecast over Taiwan with Day-1, 5, 10, and 15 forecast lead times for the period 2000-2019.

4. Summary and conclusions:

Temperature is a key weather component that affects many sectors, such as energy, aviation, communication, and agriculture. According to the IPCC report (2013), the global mean temperature has increased by 0.13°C per decade in the last 50 years, which is double the rate of the previous century. Heat waves have become more frequent and intense in recent decades, with notable occurrences in Asia. In Taiwan, the air temperature has risen by 1.4°C from 1911 to 2005, which is twice the increase in the Northern Hemisphere. It is evident that a 1°C increase in surface air temperature above 31.5°C in Taipei increases respiratory mortality by 9.3%, while a 1°C increase above 25.2°C increases cardiovascular mortality by 1.1%. To better predict and prepare for extreme weather events, extended-range or sub-seasonal weather forecasts are used to predict these heatwaves in advance. However, there is still a gap in current numerical prediction systems for the sub-seasonal extended range time scale of 10 days to one month. This is mainly due to the complexity of tropical processes, which are influenced by interactions between ocean-land-atmosphere, atmospheric circulation, convection, cloud and radiation, precipitation, and moisture on different spatial and temporal scales. To improve prediction skills on this time scale in small regions such as

Taiwan Island, global models need to be improved to better represent land-sea contrast and topography. In addition to improving the global models, post-processing techniques are also essential.

In September 2020, NOAA NCEP upgraded its Global Ensemble Forecast System to version 12 (GEFSv12) to improve the accuracy of sub-seasonal forecasts for meteorological and hydrological applications. This model was used to generate consistent reforecast products based on daily 00 UTC initial conditions for forecasts extended up to 16 days with 5 ensemble members for a period of 2000-2019, except every Wednesday when the forecasts were integrated up to 35 days with 11 members. The output of the model is subject to a high degree of uncertainty and is rarely used as-is. Therefore, post-processing techniques are used to reduce the uncertainty and improve the accuracy of the forecasts. In this study, a Hybrid calibration method combining Artificial Neural Network (ANN) and Quantile-Quantile mapping (QQ) techniques were applied to the GEFSv12 reforecasts to enhance the accuracy of summer daily T_{max} and related T_{max} extremes over Taiwan. The performance of the Hybrid technique was evaluated against ERA5 reanalysis and compared to the Raw, ANN, and QQ techniques using standard skill metrics for deterministic and ensemble probabilistic forecasts.

The GEFSv12 model was found to accurately replicate the spatial patterns of maximum temperature and its variability in Taiwan for all forecast lead times. However, it had a warm bias and overestimated the Interannual Variability (IAV) of T_{max} in the southern and inland regions of Taiwan. The RMSE of the raw model and all three calibration methods increased with increasing forecast lead time. The Raw forecast for summer T_{max} over Taiwan had a surprisingly high RMSE for all forecast lead times, ranging from However, all the calibration methods such as QQ (0.8-1.2℃), ANN (0.6-1℃), and Hybrid (0.6-1℃) notably reduced the RMSE for all forecast lead times. The QQ method yielded the highest RMSE compared to the ANN and Hybrid methods for all forecast lead times. The RMSE from the ANN and Hybrid methods were similar for all forecast lead times. Calibration techniques were effective in reducing the warm bias of ~ 0 ℃ in Taiwan during summer for all forecast lead times. The GEFSv12 model shows a strong correlation with summer daily T_{max} over Taiwan, with a coefficient of more than 0.8 for Day-1 lead time forecasts. This correlation decreases with increasing forecast lead time, ranging from 0.8 to 0.4. No improvement was observed in the correlation coefficient when using the QQ method compared to the Raw products for all forecast lead times. However, the ANN and Hybrid calibration methods showed a significant improvement in the correlation coefficient for all forecast lead times, with the improvement being more pronounced for longer lead time forecasts. The IOA of GEFSv12 for predicting Taiwan's T_{max} shows a decrease from 0.8 to 0.6 as the forecast lead time increases, with higher values for shorter lead times. All calibration methods demonstrated a significant increase in the IOA of predicting daily summer T_{max} over Taiwan for all forecast lead times. The ANN and Hybrid methods achieved scores of 0.88-0.92, while the QQ method had scores of 0.67-0.9. The Hybrid method yielded higher IOA values than the ANN for all forecast lead times.

The GEFSv12 model overestimated the number of Heatwave days over Taiwan, but this was reduced by the QQ, ANN, and Hybrid methods. The ANN model had the lowest number of heatwave days compared to the QQ and Hybrid approaches. The Hybrid method had the highest statistical categorical skill scores for all forecast lead times, outperforming the other two methods in terms of ETS, TS, SR, ACC, FAR, POD, and Frequency Bias. The prediction accuracy of Raw and all calibration methods for summer daily T_{max} extremes over Taiwan is higher for Week-1, Week-2, and Week-1 to Week-2 forecasts than for day-to-day forecasts. The comparison of Week-1, 2, and 1 to 2 from Raw and all three calibration methods reveals that the prediction skill of Week-1 summer T_{max} extreme days is superior to that of Week-2 and Week-1 to 2 when using Raw and all three calibration methods. The Hybrid method is the most effective than the other two methods. The Hybrid method of forecasting extreme T_{max} over Taiwan was found to be more effective than either QQ or ANN alone, based on the evaluation of probabilistic skill scores (Reliability, Resolution, Brier score, Brier skill score, and ROC curve). The Hybrid-calibrated GEFSv12 forecast can be beneficial in managing climate risk in Taiwan by providing extended-range forecasts of T_{max} and associated extremes.

References

1. IPCC, 2013 Climate Change **2013.** The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change ed T F Stocker, D Qin, G-K Plattner, M Tignor, S K Allen, J Boschung, A Nauels, Y Xia, V Bex and P M Midgley (Cambridge: Cambridge University Press) Summary for policymakers

2. Honda, Y. K.; Sugimoto; Ono, M. Adaptation to climate change at population level in Japan. Epidemiology **2011;** 22(1), S26. https://doi.org/10.1097/01.ede.0000391738.93182.aa

3. Nageswararao M. M.; Mohanty, U. C.; Kiran Prasad, S.; Krishna K. Osuri; Ramakrishna, S. S. V. S. Performance evaluation of NCEP climate forecast system for the prediction of winter temperatures over India. Theor. Appl. Climatol. **2016;** 126(3-4), 1–15, https://doi.org/10.1007/s00704-015-1588-6, ISSN: 0177-798X

4. Nageswararao M.M.; Sinha, P.; Mohanty, U.C.; Mishra, S. Occurrence of More Heat Waves Over the Central East Coast of India in the Recent Warming Era. Pure Appl. Geophys. **2020;** 177(2), 1143–1155, DOI: https://doi.org/10.1007/s00024-019-02304-2

5. Karrevula, N.R.; Dandi A. Ramu; Nageswararao, M. M.; Suryachandra Rao, A.; Inter-annual variability of pre-monsoon surface air temperatures over India using the North American Multi-Model Ensemble models during the global warming era. Theor. Appl. Climatol. **2023;** 151,133–151, https://doi.org/10.1007/s00704-022-04269-0

6. Rastogi, D.; Lehner, F.; Ashfaq, M. Revisiting recent U.S. heat waves in a warmer and more humid climate. Geophys. Res. Lett., **2020;** 47, https://doi.org/10.1029/2019GL086736.

7. Fischer, E. M.; Schar, C. Consistent geographical patterns of changes in high-impact European heatwaves. Nat. Geosci. **2010;** 3, 398–403. https://doi.org/10.1038/NGEO866.

8. Meehl, G. A.; Tebaldi, C. More intense, more frequent, and longer lasting heat waves in the 21st century. Science. **2004;** 305. 994–7

9. Smoyer-Tomic, K.E.; Kuhn, R.; Hudson, A. Heat Wave Hazards: An Overview of Heat Wave Impacts in Canada. Nat Hazards, **2003;** 28, 465–486, https://doi.org/10.1023/A:1022946528157

10. Lin, C.Y.; Chua, Y.J.; Sheng, Y.F.; Hsu, H.H.; Cheng, C.T.; Lin, Y.Y. Altitudinal and latitudinal dependence of future warming in Taiwan simulated by WRF nested with ECHAM5/MPIOM. Int. J. Climatol., **2015;** 35, 1800–1809. https://doi.org/10.1002/joc.4118

11. IPCC, 2007 Intergovernmental panel on climate change, Fourth Assesment Report: Climate Change **2007**. Geneva.

12. Chung, J.Y.; Honda, Y.; Hong, Y.C.; Pan, X.C.; Guo, Y.L.; Kim, H. Ambient temperature and mortality: an international study in four capital cities of East Asia. Sci Total Environ **2009;** 408, 390–396

13. Mariotti, A.; Ruti, P.M.; Rixen, M. Progress in subseasonal to seasonal prediction through a joint weather and climate community effort. npj Clim. Atmos. Sci., **2018;** 1, 4. https://doi.org/10.1038/s41612-018-0014-z

14. Li, S.; Roberston, A. W. ; Li, S.; Robertson, A. W. Evaluation of submonthly precipitation forecast skill from global ensemble prediction systems. Mon. Wea. Rev., **2015;** 143, 2871–2889, https://doi.org/10.1175/MWR-D-14-00277.1

15. Vitart, F; et al. The sub-seasonal to seasonal prediction project (S2S) and the prediction of extreme events. Bull Am Meteorol Soc**., 2017;** 98, 163-173, https://doi.org/10.1175/BAMS-D-16-0017.1

16. Nageswararao M. M.; Yuejian Zhu; Vijay Tallapragada. Prediction Skill of GEFSv12 for Southwest Summer Monsoon Rainfall and Associated Extreme Rainfall Events on Extended Range scale over India. Wea. Forecasting, **2022a;** 37(7), 1135–1156, https://doi.org/10.1175/WAF-D-21-0184.1

17. Nageswararao M. M.; Yuejian Zhu; Vijay Tallapragada; Meng-Shih Chen. Prediction Skill of GEFSv12 in Depicting Monthly Rainfall and Associated Extreme Events over Taiwan during the Summer Monsoon. Wea. Forecasting, **2022b;** 37(12), 2239–2262, https://doi.org/10.1175/WAF-D-22-0025.1

18. Kang, I.-S.; et al. Intercomparison of the climatological variations of Asian summer monsoon precipitation simulated by 10 GCMs. Clim. Dyn., **2022;** 19, 383–395, https://doi.org/10.1007/s00382-002-0245-9

19. Glahn, H. R.; Lowry, D. A. The use of model output statistics (MOS) in objective weather forecasting. J. Appl. Meteorol., **1972;** 11, 1203, https://doi.org/10.1175/1520-0450(1972)011%3C1203:TUOMOS%3E2.0.CO;2

20. Hamill, T.; Colucci, S. Evaluation of Eta-RSM ensemble probabilistic precipitation forecasts. Mon. Wea. Rev., **1998;** 126, 711–724.

21. Mert, G.; Yang, D.; Srinivasan, D. Ensemble solar forecasting using data-driven models with probabilistic post-processing through GAMLSS. Sol. Energy, **2020;** 208, 612-622, 10.1016/j.solener.2020.07.040

22. Li M.; Wang, Q.J.; Robertson, D.E.; Bennett, J.C. Improved error modelling for streamflow forecasting at hourly time steps by splitting hydrographs into rising and falling limbs. J. Hydrol., **2017;** 555 , 586-599, 10.1016/j.jhydrol.2017.10.057

23. Wilks, D. S. Chapter 3: Univariate ensemble forecasting S. Vannitsem, D.S. Wilks, J.W. Messner (Eds.), Statistical Postprocessing of Ensemble Forecasts**, 2018**; pp. 49-89, 10.1016/C2016-0-03244-8

24. Ebert, E. E. Ability of a poor man's ensemble to predict the probability and distribution of precipitation. Mon. Wea. Rev., **2001;** 129, 2461-2480, https://doi.org/10.1175/1520-0493(2001)129%3C2461:AOAPMS%3E2.0.CO;2 .

25. Hamill, T. M.; Whitaker, J. S. Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. Mon. Wea. Rev., **2006;** 134, 3209-3229, https://doi.org/10.1175/MWR3237.1 .

26. Zhu, Y.; Luo, Y. Precipitation calibration based on frequency matching method (FMM). Wea. Forecasting, **2015**; 30, 1109–1124, https://doi.org/10.1175/WAF-D-13-00049.1.

27. Guan, H.; et al. GEFSv12 reforecast dataset for supporting subseasonal and hydrometeorological applications. Mon. Wea. Rev., **2022;** 150, 647–665, https://doi.org/10.1175/MWR-D-21-0245.1.

28. Zhao, T., J.; Bennett, C.; Wang, Q.; Schepen, A. ; Wood, A. W.; Robertson, D. E.; Ramos, M.-H. How suitable is quantile mapping for post processing GCM precipitation forecasts?. J. Climate, **2017;** 30, 3185–3196, https://doi.org/10.1175/JCLI-D-16-0652.1.

29. Verkade, J.S.; Brown, J. D.; Reggiani, P.; Weerts, A.H. Post-processing ECMWF precipitation and temperature ensemble reforecasts for operational hydrologic forecasting at various spatial scales. J. Hydrol., **2013;** 501 , 73-91, 10.1016/j.jhydrol.2013.07.039

30. Wang, Q. J.; Shao, Y.; Song, Y. ; Schepen, A.; Robertson, D.E.; Ryu, D.; Pappenberger, F. An evaluation of ECMWF SEAS5 seasonal climate forecasts for Australia using a new forecast calibration algorithm. Environ. Model. Softw., **2019;** 122, 104550, https://doi.org/10.1016/j.envsoft.2019.104550

31.   Zhou, X.; Zhu, Y.; Fu, B.; Hou, D.; Peng, J.; Luo, Y.; Li, W. The development of the Next NCEP Global Ensemble Forecast System. Science and Technology Infusion Climate Bulletin, NOAA's National Weather Service, 43rd NOAA Annual Climate Diagnostics and Prediction Workshop (CDPW), **2019,** 159-163.

32.   Zhou, X.; Zhu, Y.; Hou, D.; Fu, B.; Li, W.; Guan, H.; Sinsky, E.; Kolczynski, W.; Xue, X.; Luo, Y.; et al. The Development of the NCEP Global Ensemble Forecast System Version 12. Weather Forecast**., 2022**; 37, 727, https://doi.org/10.1175/WAF-D-21-0112.1

33.   Hamill, T. M.; et al. The Reanalysis for the Global Ensemble Forecast System, version 12. Mon. Wea. Rev., **2022**; 150(1), 59–79, https://doi.org/10.1175/MWR-D-21-0023.1

34.   Harris, L. M.; Lin, S.-J. A two-way nested global-regional dynamical core on the cubed-sphere grid. Mon. Wea. Rev., **2013;** 141(1), 283-306, https://doi.org/10.1175/MWR-D-11-00201.1.

35.   Han, J.; Wang, W.; Kwon, Y. C.; Hong, S. -Y.; Tallapragada, V.; Yang, F. Updates in the NCEP GFS cumulus convection schemes with scale and aerosol awareness. Wea. Forecasting, **2017;** 32, 2005-2017, https://doi.org/10.1175/WAF-D-17-0046.1 .

36.   Han, J.-Y.; Hong, S.-Y.; Lim Sunny, K.-S.; Han, J. Sensitivity of a cumulus parameterization scheme to precipitation production representation and its impact on a heavy rain event over Korea. Mon. Wea. Rev., **2016;** 144, 2125-2135, https://doi.org/10.1175/MWR-D-15-0255.1 .

37.   Clough, S. A.;   Shephard, M. W.; Mlawer, E. J.; Delamere, J. S.; Iacono, M. J. ; Cady-Pereira, K.; Boukabara, S.; Brown, P. D. Atmospheric radiative transfer modeling: A summary of the AER   codes. J.   Quant.   Spectrosc.   Radiat.   Transfer**, 2005;**   91, 233–244, https://doi.org/10.1016/j.jqsrt.2004.05.058.

38.   Chun, H.-Y.;   Baik, J.-J. Momentum flux by thermally induced internal gravity waves and its approximation for large-scale models. J. Atmos. Sci., **1998;** 55, 3299–3310, https://doi.org/10.1175/1520-0469(1998)055<3299:MFBTII>2.0.CO;2.

39.   Alpert, J. C.;   Kanamitsu, M.;   Caplan, P. M.;   Sela, J. G.;   White, G. H. ; Kalnay, E.   Mountain induced gravity wave drag parameterization in the NMC medium-range forecast model. Eighth Conf. on Numerical Weather Prediction, **1988;** Baltimore, MD, Amer. Meteor. Soc., 726–733.

40.   Zhu, Y.; Zhou, X.; Pena, M.; Li, W.; Melhauser, C.; Hou, D. Impact of sea surface temperature forcing on weeks 3 and 4 forecast skill in the NCEP Global Ensemble Forecast System. Wea. Forecasting, **2017;** 32, 2159-2173, https://doi.org/10.1175/WAF-D-17-0093.1.

41.   Zhu, Y.; Zhou, X.; Wei Li; Dingchen, H.; Christopher, M.; Eric Sinsky, Malaquias, P.; Bing Fu; Hong, G.; Walter, K.; Richard, W.; Tallapragada, V.; Toward the improvement of sub-seasonal prediction in the NCEP Global Ensemble Forecast System (GEFS). J. Geophys. Res.: Atmos., **2018;** 123, 6732-6745, https://doi.org/10.1029/2018JD028506

42.   Li, W.; ;et al. Evaluating the MJO forecast skill from different configurations of NCEP GEFS extended forecast. Climate Dyn., **2019;** 52, 4923–4936, https://doi.org/10.1007/s00382-018-4423-9.

43.   Shutts, G. A kinetic energy backscatter algorithm for use in ensemble prediction systems. Q.J. Roy. Meteor. Soc., **2005;** 131, 3079-3102, https://doi.org/10.1256/qj.04.106 .

44.   Shutts, G.; Palmer, T. N. The use of high-resolution numerical simulations of tropical circulation to calibrate stochastic physics schemes. Proc. Workshop on Simulation and Prediction of Intra-Seasonal Variability with   Emphasis   on   the   MJO.   Reading,   United   Kingdom,   ECMWF,   **2004;**   83-102, https://www.ecmwf.int/node/12212 .

45.   Buizza, R.; Miller, M.; Palmer, T. Stochastic representation of model uncertainties in the ECMWF Ensemble Prediction System. Quart. J. Roy. Meteor. Soc**., 1999;** 125, 2887-2908, https://doi.org/10.1002/qj.49712556006.

46.   Palmer, T. N.; Buizza, R.; Doblas-Reyes, F.; Jung, T.; Leutbecher, M.; Shutts, G. ; Steinheimer, M.; Weisheimer, A. Stochastic parametrization and model uncertainty. ECMWF Tech. Memo. , **2009;** 598, 42, https://www.ecmwf.int/node/11577

47.   Hersbach, H.; Bell, B.; Berrisford, P.; Biavati, G.; Horányi, A. ; Muñoz Sabater, J.; Nicolas, J.; Peubey, C.; Radu, R.; Rozum, I.; Schepers, D.; Simmons, A.; Soci, C.; Dee, D.; Thépaut, J-N. ERA5 hourly data on single levels from 1959 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS), **2018;**10.24381/cds.adbb2d47

48.   Piani, C.; Haerter, J.; Coppola, E. Statistical bias correction for daily precipitation in regional climate models over Europe. Theor. Appl. Climatol., **2010;** 99(1-2), 187-192, https://doi.org/10.1007/s00704-009-0134-9

49.   Saxena, A.; Verma, N.; Tripathi, K. C. A review study of weather forecasting using artificial neural network approach. Int J Eng Res Technol., **2013;** 2(11), 2029–2035.

50. Bani-ahmad, S.; Alshaer, J.; Al-oqily, I. Development of temperature-based weather forecasting models using neural networks and fuzzy logic. Int J Multimed Ubiquitous Eng., 2014; 9(12), 343–366, http://dx.doi.org/10.14257/ijmue.2014.9.12.31.

51. Feng, J.; Lu, S. Performance analysis of various activation functions in artificial neural networks. J Phys: Conf Ser., **2019;** 1237, 022030, 10.1088/1742-6596/1237/2/022030

52. Abbot, J.; Marohasy, J. Input selection and optimisation for monthly rainfall forecasting in Queensland, Australia, using artificial neural networks. Atmos. Res., **2014;** 138,166-178, https://doi.org/10.1016/j.atmosres.2013.11.002.

53. Yilmaz, A. G.; Imteaz, M.A.; Jenkins, G. Catchment flow estimation using artificial neural networks in the mountainous Euphrates Basin. J Hydrol , **2011;** 410, 134–140, https://doi.org/10.1016/j.jhydrol.2011.09.031

54. Ahmad, R.; Lazin, N. M.; Samsuri, S. F. M. Neural network modeling and identification of naturally ventilated tropical greenhouse climates. Wseas. Trans. Syst. Control., 2014; 9, 445-453.

55. McCulloch; Warren; Walter Pitts. A Logical Calculus of Ideas Immanent in Nervous Activity. Bulletin of Mathematical Biophysics. **1943**; 5 (4), 115–133, doi:10.1007/BF02478259.

56. Fausett, L. Fundamentals of Neural Network. Prentice Hall, Hoboken, **1994**.

57. Singh, G.; Panda, R. K. Daily sediment yield modeling with artificial neural network using 10-fold cross validation method: a small agricultural watershed, Kapgari, India. International Journal of Earth Sciences and Engineering., **2011**; 4, 443–450.

58. Singh, G.; Panda, R. K. Bootstrap-based artificial neural network analysis for estimation of daily sediment yield from a small agricultural watershed. International Journal of Hydrology Science and Technology., **2015**; 5, 333–348. doi:10.1504/IJHST.2015.072634

59. Hecht-Nielsen, R. Kolmogorov's mapping neural network existence theorem. Proceedings of the international conference on Neural Networks (Vol. 3, pp. 11- 14). New York, NY, USA: IEEE Press.; **1987**.

60. Nair, A.; Singh, G.; Mohanty, U.C. Prediction of Monthly Summer Monsoon Rainfall Using Global Climate Models Through Artificial Neural Network Technique. Pure Appl. Geophys., **2018;** 175, 403–419, https://doi-org.proxy-um.researchport.umd.edu/10.1007/s00024-017-1652-5

61. Johnstone, C.; Sulungu, E. D. Application of neural network in prediction of temperature: a review. Neural Comput & Applic., **2021**; 33, 11487–11498, https://doi-org.proxy-um.researchport.umd.edu/10.1007/s00521-020-05582-3

62. Roebber, P. J. Visualizing multiple measures of forecast quality. Wea. Forecasting, **2009**; 24, 601–608, https://doi.org/10.1175/ 2008WAF2222159.1

63. Brier, G. W. Verification of forecasts expressed in terms of probability. Mon. Wea. Rev, **1950**; 78, 1-3, https://doi.org/10.1175/1520-0493(1950)078%3C0001:VOFEIT%3E2.0.CO;2 .

64. Toth, Z.; Talagrand, O.; Candille, G.; Y. Zhu. Probability and ensemble forecasts. Forecast Verification: A Practitioner's Guide in Atmospheric Science, **2003**; I. T. Jolliffe and D. B. Stephenson, Eds., Wiley, 137-63.

65. Weijs, S. V.; Van Nooijen, R.; van de Giesen, N. Kullback–Leibler divergence as a forecast skill score with classic reliability–resolution–uncertainty decomposition. Monthly Weath. Rev., **2010**; 138, 3387–3399, https://doi.org/10.1175/2010MWR3229.1

66. Marzban, C. The ROC curve and the area under it as performance measures. Weather Forecast., **2004**; 19(6),1106–1114, https://doi.org/10.1175/825.1