

# CALIBRATING BAYESIAN UNET++ FOR SUB-SEASONAL FORECASTING

Büşra Asan<sup>1</sup>\*, Abdullah Akgül<sup>2</sup>, Alper Ünal<sup>3</sup>, Melih Kandemir<sup>2</sup>, and Gözde Ünal<sup>1</sup>

<sup>1</sup>Department of Computer Engineering, Istanbul Technical University

<sup>2</sup>Department of Mathematics and Computer Science, University of Southern Denmark

<sup>3</sup>Eurasia Institute of Earth Sciences, Istanbul Technical University

## ABSTRACT

Seasonal forecasting is a crucial task when it comes to detecting the extreme heat and colds that occur due to climate change. Confidence in the predictions should be reliable since a small increase in the temperatures in a year has a big impact on the world. Calibration of the neural networks provides a way to ensure our confidence in the predictions. However, calibrating regression models is an under-researched topic, especially in forecasters. We calibrate a UNet++ based architecture, which was shown to outperform physics-based models in temperature anomalies. We show that with a slight trade-off between prediction error and calibration error, it is possible to get more reliable and sharper forecasts. We believe that calibration should be an important part of safety-critical machine learning applications such as weather forecasters.

## 1 INTRODUCTION

Seasonal forecasting is a crucial task when it comes to foreseeing the effects of climate change, especially in making predictions and decisions based on these effects. Generating accurate seasonal and sub-seasonal forecasts demands substantial resources, such as the curation of Coupled Model Intercomparison Projects (CMIP) datasets. These datasets combine outputs from over a hundred climate models worldwide, facilitating top-tier climate simulations. Leveraging the vast data reservoirs from CMIP6 (Eyring et al., 2016), the latest phase of CMIP, there are ongoing efforts to harness deep learning methodologies for enhanced climate forecasting. For instance, Luo et al. (2022) use Bayesian Neural Networks (BNN) with CMIP6 for climate prediction in the North Atlantic, and Anochi et al. (2021) use CMIP6 to assess precipitation. Andersson et al. (2021) forecast the change in Arctic sea ice area with the same dataset. In this work, we also utilize the CMIP6 dataset to produce well-calibrated and sharp forecasts which are crucial for climate sciences (Gneiting et al., 2007).

We expand the capabilities of the forecast model introduced in Ünal et al. (2023) using the calibration approach proposed by Kuleshov et al. (2018). This model is shown to achieve better performance than physics-based methods on sub-seasonal forecasting, especially at predicting temperature anomalies that indicate extreme hot and cold temperatures that are crucial for climate change.

In this work, we calibrate the Bayesian version of the forecaster in a regression setting. We show that the BNNs produce the most calibrated and sharp forecasts. We compare the performance of BNNs, Monte Carlo Dropout (MC-Dropout) Gal & Ghahramani (2016), and Deep Ensemble Lakshminarayanan et al. (2017) methods for assessing climate forecast uncertainty and sharpness, exploring the potential for improved reliability through calibration. Our contributions can be listed as follows:

1. We apply calibration to a sub-seasonal forecaster that is able to predict extreme events better than simulations. We show that calibrating deep learning models should be a crucial step while applying deep learning to climate sciences.

---

\*Corresponding author: asan18@itu.edu.tr

2. We show that well-calibrated forecasters not only produce better confidence intervals but may also improve the sharpness of the forecasts.
3. This method may be generalized to any other application in climate sciences that gives critical importance to the reliability of the results such as extreme events, precipitation, and natural disasters such as earthquakes, floods, and drought.

## 2 METHODOLOGY

We formulate the problem as predicting the monthly average air temperature at 2 meters above the earth surface ( $2m$ ) for each coordinate in a 2D temperature grid which we will name as the temperature map. Our aim is to construct a reliable confidence interval for each coordinate since Bayesian methods often produce uncalibrated results Kuleshov et al. (2018).

We first train the model on CMIP6 climate simulations, then fine-tune it with ERA5 reanalysis data based on real climate measurements. We denote the 2D temperature map at time  $t$  as  $x_t$ . Train set  $D = \{X_t, Y_t\}_{t=1}^T$  consists of stacked monthly time-series temperature maps as the input. The input  $X_t$  refers to  $x_{t-1:t-k-m}$  which denotes the range of the stacked months and  $Y_t$  corresponds to  $x_t$ . The periodical month selection process from the given range is described in Unal et al. (2023) and the same setting is used to make a fair comparison.

### 2.1 BAYESIAN UNET++ FOR TEMPERATURE PREDICTION

For temperature forecasting, we convert UNet++ into a BNN (Goan & Fookes, 2020). BNNs are highly regarded for their capability in quantifying uncertainty, offering robust insights into predictive models Kristiadi et al. (2020). Thus, we converted the final three layers of the UNet++ architecture into Bayesian convolutional layers, where we model the weights  $\theta$  of the neural network as a Gaussian distribution. Letter, we maximize evidence lower bound Kingma & Welling (2022).

### 2.2 UNCERTAINTY ANALYSIS

**Confidence Intervals** are used for measuring the uncertainty. Quantiles are calculated from the predictions and checked whether the correct portion of the predictions actually conforms to those intervals.

**Calibration** in neural networks refers that if the confidence interval is chosen as 95%, then the intervals should capture around 95% of the observed outcomes  $Y_t$ . To measure calibration, we count observations that stay below the predicted upper bound for the quantile  $p$  of the sample  $t$ , then normalize with the size of the dataset. A neural network is said to be calibrated if it satisfies the following

$$\frac{1}{T} \sum_{t=1}^T (\mathbb{1}\{Y_t \leq F_t^{-1}(p)\}) \longrightarrow p, \quad (1)$$

as  $T \rightarrow \infty$  (Gneiting et al., 2007), where  $F_t$  refers to the Cumulative Distribution Function (CDF) of the output of the neural network  $H$  for the input  $X_t$ .

**Calibrated Regression.** We need to match empirical and predicted CDFs according to Equation 1. Therefore, the training partition of the ERA5 dataset is used to construct a calibration dataset to map the predicted CDF to the empirical CDF. We train a new regressor  $R : [0, 1] \rightarrow [0, 1]$  on the calibration dataset. Thus, we expect  $R \circ F_k$  to be calibrated. CDFs are monotonically increasing functions, hence the choice of  $R$  is an Isotonic Regressor (Niculescu-Mizil & Caruana, 2005).

From the training partition  $S = \{X_{t-1:t-k-m}, Y_t\}_{t=1}^{T'}$  of ERA5, calibration dataset  $C = \{c_t, y_t\}_{t=1}^{T'}$  is constructed where  $c_t$  refers to  $F_t(Y_t)$  and  $y_t$  refers to  $\hat{P}(F_t(Y_t))$  using the dataset generation method in Kuleshov et al. (2018).  $\hat{P}$  is formulated as

$$\hat{P} = \frac{1}{T} |\{Y_t | F_t(Y_t) < p, t = 1, \dots, T\}| \quad (2)$$

where  $|A|$  refers to the cardinality of the set A. It calculates empirical CDF from the predicted CDF by normalizing the count of output  $Y_t$  staying below  $p^{th}$  quantile of  $F_t$ .

Table 1: Metrics (lower the better) for calibrated and uncalibrated versions of the models.

Metrics	Uncalibrated			Calibrated		
	CE	MAE	Sharpness	CE	MAE	Sharpness
UNet++	N/A	<b>0.975</b>	N/A	N/A	N/A	N/A
Bayesian UNet++	<b>0.023</b>	2.237	<b>0.291</b>	<b>0.015</b> ( $\downarrow 34.8\%$ )	2.298 ( $\uparrow 2.7\%$ )	<b>0.274</b> ( $\downarrow 6.9\%$ )
Dropout (40%)	0.131	0.993	0.853	0.035 ( $\downarrow 73.2\%$ )	<b>0.990</b> ( $\downarrow 0.3\%$ )	0.847 ( $\downarrow 0.7\%$ )
Deep Ensemble	0.086	1.548	0.789	0.024 ( $\downarrow 70.0\%$ )	1.366 ( $\downarrow 11.8\%$ )	0.799 ( $\uparrow 1.3\%$ )

As a result, the estimated  $\mathbb{P}(Y \leq F_X^{-1}(p))$  by the regressor  $R$  provides the calibrated probability that a random  $Y$  falls into the credible interval so that we can adjust the predicted probability to the empirical probability.

### 2.3 TRAINING & EVALUATION

We train our model using the  $2m$  temperature variable from 9 ensembles of the CMIP6 dataset. 1700 samples are separated for training and 100 for validation. 400 samples from the ERA5 dataset are used for fine-tuning and the construction of the calibration dataset. 116 samples from ERA5 are used in the evaluation of all methods as in Unal et al. (2023).

**Metrics.** To measure accuracy, Mean Absolute Error (MAE) is used. MAE for calibrated models in Table 1 is recalculated using the mid-quantile values from the calibrated forecaster.

Sharpness (Gneiting et al., 2007) is a metric which is widely used in climate forecasting. It measures the concentration of the forecasts as

$$\text{sharpness}(F_1, F_2, \dots, F_T) = \frac{1}{T} \sum_{t=1}^T \text{var}(F_t). \quad (3)$$

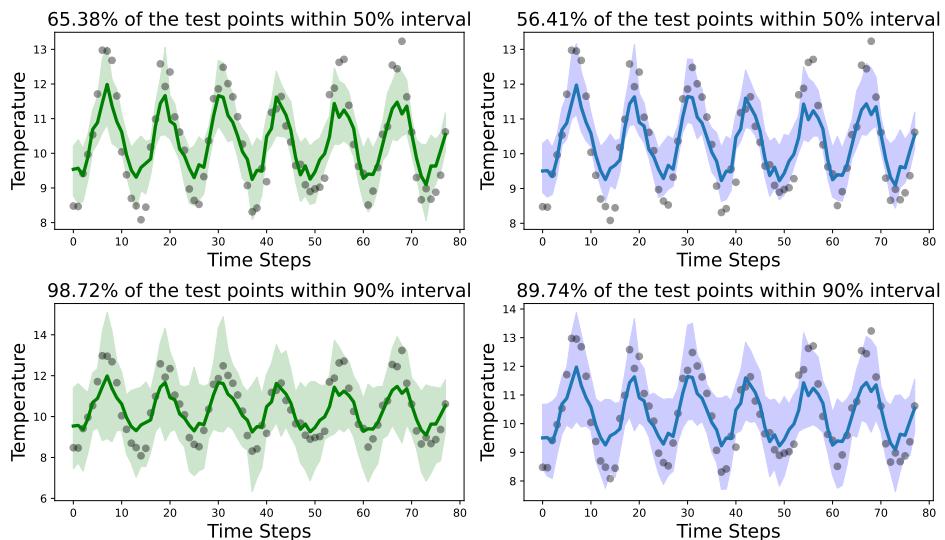


Figure 1: 50% confidence interval (Top) and 90% confidence interval (Bottom) of the Bayesian UNet++ for a sample in the North West Coast of America are given. The mean coverage percentages for confidence intervals are 63% and 91% for the calibrated, and 66% and 98% for the uncalibrated models. Thus, we choose a representative sample. Uncalibrated confidence interval plots are shown on the left (green), and calibrated plots are on the right (blue). Grey dots refer to the average temperature values for each month in the given time period (2016-2021). The percentage of the values falling within the intervals aligns more closely with the expected confidence levels, both at 50% and 90% in the calibrated model's plot.

Calibration error (CE) proposed by Kuleshov et al. (2018) is used for assessing the quality of the calibration of the forecasts as

$$\text{CE}(F_1, Y_1, \dots, F_T, Y_T) = \frac{1}{m} \sum_{i=1}^m w_j(p_j - \hat{p}_j) \quad (4)$$

where  $m$  refers to the number of confidence levels  $0 \leq p_1 \leq \dots \leq p_j \leq \dots \leq p_m \leq 1$  and  $\hat{p}_j$  is empirical frequency. In this setting,  $w_j$  is chosen as 1.

### 3 RESULTS

We use the experimental settings of Unal et al. (2023). Table 1 illustrates the impact of calibration on uncertainty quantification methods. The Bayesian model demonstrates the highest sharpness and calibration as we expected. However, there exists a trade-off between MAE and CE, with Bayesian demonstrating the lowest CE, followed by Ensemble and Dropout. Apart from the reduction in CE, we observe a decrease in MAE for Dropout and Deep Ensemble which suggests that calibration not only improves the accuracy of the network performance but also enhances the capture probability percentages of confidence intervals around point estimates. MAE is calculated using the actual 50% quantile values predicted by the Isotonic Regressor.

Figure 1 demonstrates that for the calibrated case, roughly 90% of the 90% confidence intervals capture the true temperature values in the test dataset. We also observe the same result for 50% interval. Thus, the proposed calibration produced results in line with the expected proportion of confidence intervals capturing the true outcome at the given confidence level, suggesting that the model is well-calibrated.

CE is visualized in Figure 2. Equation 4 is applied to the values calculated for the calibration plot for each quantile, and the mean is used as the CE of that sample. After the calibration, our model converges to the  $y = x$  line which indicates that the predicted confidences for the samples are closer to expected confidences, especially for quantiles larger than 50%.

### 4 CONCLUSION

We proposed a method to enhance the sharpness and reliability of weather forecasts by calibrating them using a CDF-based calibration approach. This involved transforming the final layers of UNet++ to Bayesian. Periodically stacked multi-dimensional time-series data used as input. As we designed the output of the network to produce a CDF, we trained an isotonic regressor to calibrate the confidence intervals. We benchmarked the calibrated and uncalibrated results of three uncertainty quantification methods. Furthermore, we show that calibrating Dropout and Deep Ensemble might increase the accuracy of the network along with improving the uncertainty quantification.

This work emphasizes the significance of calibrating neural networks while suggesting potential improvements for forecast reliability. Various fields in climate sciences can benefit from calibration since uncertainties arise from incomplete modeling of the earth and the inherent complexity of climate systems. While our focus was on temperature forecasting, this approach can be extended to predicting other essential climate variables such as precipitation, pressure, and wind components.

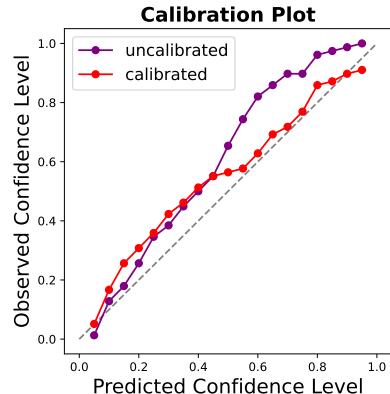


Figure 2: Calibration plot suggested by Kuleshov et al. (2018) given for a sample in the grid in Figure 1 to evaluate the calibration of the forecasts. Each predicted confidence level is plotted against its corresponding expected confidence level. Predictions illustrate the frequency of observing an outcome  $Y_t$  at each level. We expect calibrated models to be closer to  $y = x$ .

## REFERENCES

Tom R Andersson, J Scott Hosking, María Pérez-Ortiz, Brooks Paige, Andrew Elliott, Chris Russell, Stephen Law, Daniel C Jones, Jeremy Wilkinson, Tony Phillips, et al. Seasonal arctic sea ice forecasting with probabilistic deep learning. *Nature Communications*, 12(1):5124, 2021.

Juliana Aparecida Anochi, Vinícius Albuquerque de Almeida, and Haroldo Fraga de Campos Velho. Machine learning for climate precipitation prediction modeling over south america. *Remote Sensing*, 13(13):2468, 2021.

V. Eyring, S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor. Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization. *Geoscientific Model Development*, 9:1937–1958, 2016.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pp. 1050–1059. PMLR, 2016.

Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268, 2007.

Ethan Goan and Clinton Fookes. Bayesian neural networks: An introduction and survey. In *Case Studies in Applied Bayesian Data Science*, pp. 45–87. Springer International Publishing, 2020. doi: 10.1007/978-3-030-42553-1\_3.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022.

Agustinus Kristiadi, Matthias Hein, and Philipp Hennig. Being bayesian, even just a bit, fixes overconfidence in relu networks, 2020.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *International Conference on Machine Learning*, pp. 2796–2804. PMLR, 2018.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 30, 2017.

Xihai Luo, Balasubramanya T Nadiga, Ji Hwan Park, Yihui Ren, Wei Xu, and Shinjae Yoo. A bayesian deep learning approach to near-term climate prediction. *Journal of Advances in Modeling Earth Systems*, 14(10):e2022MS003058, 2022.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *International Conference on Machine Learning*, pp. 625–632, 2005.

Alper Unal, Busra Asan, Ismail Sezen, Bugra Yesilkaynak, Yusuf Aydin, Mehmet Ilicak, and Gozde Unal. Climate model-driven seasonal forecasting approach with deep learning. *Environmental Data Science*, 2:e29, 2023. doi: 10.1017/eds.2023.24.