Calibrating Bayesian UNet++ for Sub-seasonal Forecasting

Büşra Asan[1], Abdullah Akgül[2], Alper Ünal[3], Melih Kandemir[2]and Gözde Ünal[1]

[1]Department of Computer Engineering, Istanbul Technical University, [2]Department of Mathematics and Computer Science, University of Southern Denmark, [3]Eurasia Institute of Earth Sciences, Istanbul Technical University

Bayesian UNet++ Architecture and Training Loss

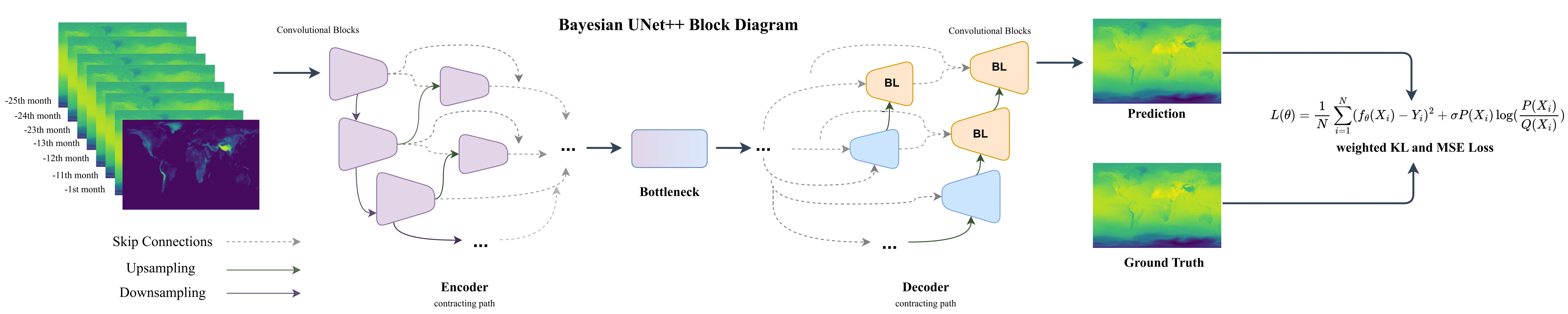

Figure 1:Architecture of Bayesian UNet++ network architecture with input-output descriptions and loss function. Orange blocks represent Bayesian convolutional layers in the network.

Introduction

We formulate the problem as predicting the monthly $(2m)$ temperature 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. We first train the model on CMIP6 climate simulations, then fine-tune it with ERA5 reanalysis data based on real climate measurements.

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.
2. We show that well-calibrated forecasters not only produce better confidence intervals but may also improve the sharpness of the forecasts.

Calibration

In neural networks, calibration 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.

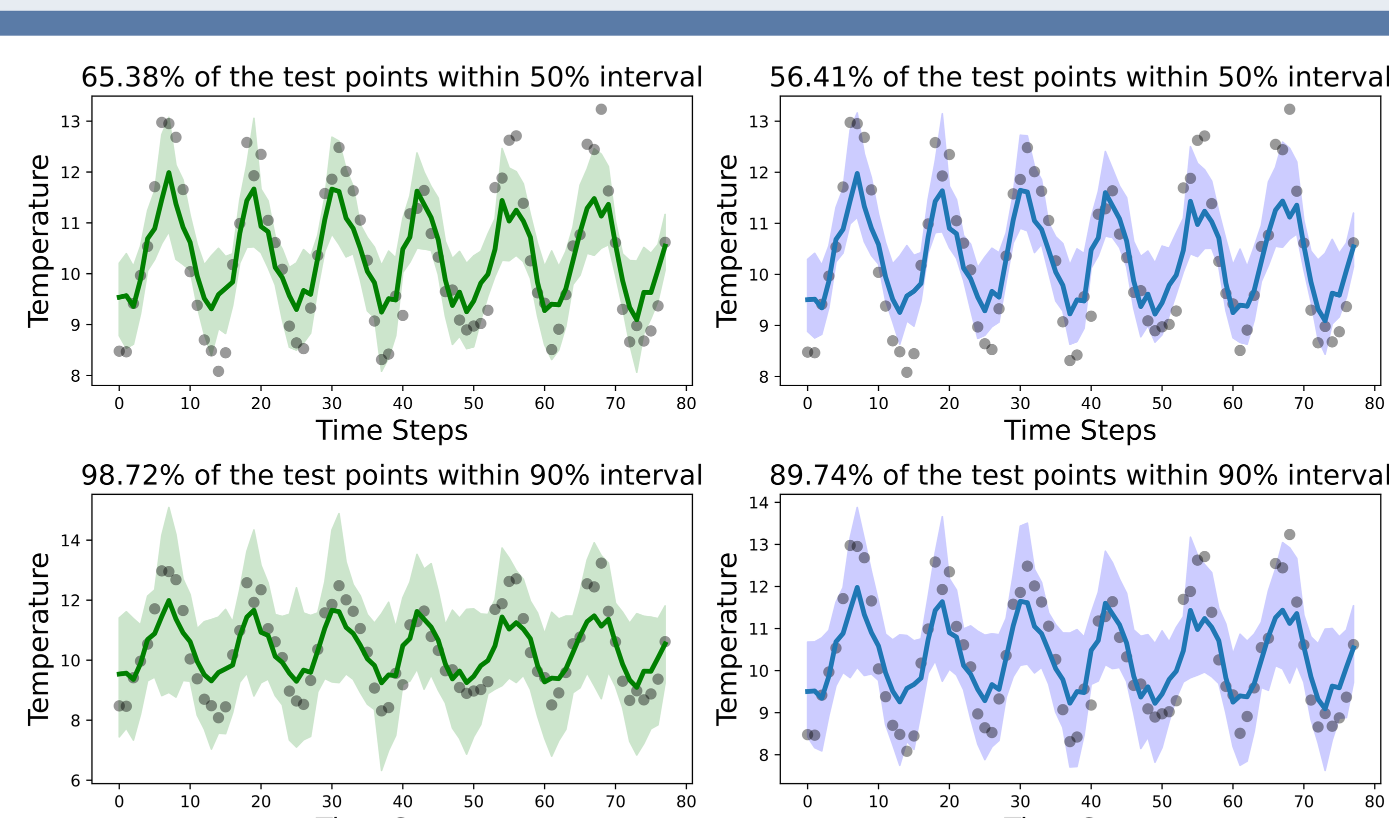

Figure 2:50% confidence interval (Top) and 90% confidence interval (Bottom) of the Bayesian UNet++ for a sample in the North West Coast of America are given.

Methodology

Calibrated Regression.

1. Train the Bayesian UNet++ on 9 ensembles of CMIP6 dataset. Then, fine-tune with ERA5. 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. F_t refers to the CDF of the UNet++ forecaster H_t.
2. 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 $\hat{P}(F_t(Y_t))$. $\hat{P}$ is
$$\hat{P} = \frac{1}{T}|\{Y_t|F_t(Y_t) < p, t = 1, ..., T\}| \quad (1)$$
It calculates empirical CDF from the predicted CDF by normalizing the count of output Y_t staying below p^{th} quantile of F_t.
3. We train an Isotonic Regressor $R : [0,1] \rightarrow [0,1]$ on the calibration dataset. Thus, we expect $R \circ F_k$ to be calibrated.
4. 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.

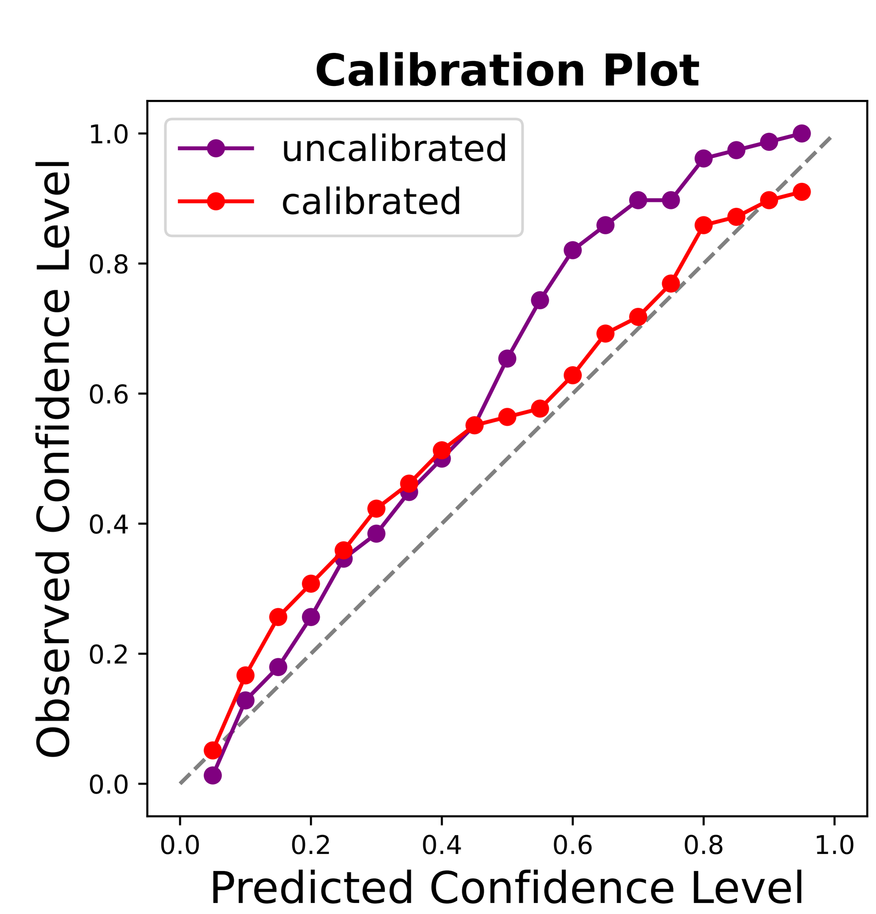

Figure 3:Calibration plot suggested by Kuleshov et al. given for a sample in the grid in Figure 2 to evaluate the calibration of the forecasts. Each predicted confidence level is plotted against its corresponding expected confidence level. Predictions illustrate the frequency of observing an outcome Y_t at each level. We expect calibrated models to be closer to $y = x$.

Results

Table 1:Metrics (lower the better) for calibrated and uncalibrated models. MAE for calibrated models is calculated using the actual 50% quantile values predicted by the Isotonic Regressor R.

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