

# Encoder-decoder-based image transformation approach for integrating precipitation forecasts

**Hiroataka Hachiya**

HHACHIYA@WAKAYAMA-U.AC.JP

*Graduate School of Systems Engineering, Wakayama University/Center for AIP, RIKEN*

**Yusuke Masumoto**

MASUMOTO.YUSUKE@G.WAKAYAMA-U.JP

*Graduate School of Systems Engineering, Wakayama University/Center for AIP, RIKEN*

**Yuki Mori**

MORI.YUKI-03@FUJITSU.COM

*Numerical Prediction Development Center, Japan Meteorological Agency/Fujitsu Limited*

**Naonori Ueda**

NAONORI.UEDA@RIKEN.JP

*Center for AIP, RIKEN*

**Editors:** Vineeth N Balasubramanian and Ivor Tsang

## Abstract

As the damage caused by heavy rainfall is becoming more serious, the improvement of precipitation forecasts is highly demanded. For this purpose, arithmetic and Bayesian average-based methods have been proposed to integrate multiple 2D-grid forecasts. However, since a single weight is shared in the entire grid in these methods, local variations of the importance of forecasts could not be taken into account. Besides, although a variety of information is available in precipitation forecast, it would not be straightforwardly to incorporate the additional information in the existing methods. To overcome these problems, we propose an encoder-decoder-based image transformation method that generates a weight image that is optimized in a pixel-wise manner, and additional information could be embedded as the channel of input images and feature maps. Through the experiment of precipitation forecast in the period from April 2018 to March 2019 in Japan, we will show that our proposed integration method outperforms existing methods.

**Keywords:** precipitation forecast, image-to-image transformation, encoder-decoder network

## 1. Introduction

Currently, Japan has been experiencing a high rainfall season for the last ten years ([Japan Meteorological Agency \(b\)](#)). Local heavy rains, e.g., cloudbursts and typhoons, cause damage every year. To reduce these damages, further improvement of precipitation forecasts is expected. Japan Meteorological Agency (JMA) has been operating three deterministic numerical weather prediction (NWP) models ([Japan Meteorological Agency \(2019\)](#)), **Global Spectral Model (GSM)**, **Meso-Scale Model (MSM)**, and **Local Forecast Model (LFM)**, each of which has a different scale of target region, resolution, and forecast time. Through statistical processes, the outputs of three NWP models are respectively converted to 2D-grid data, called **guidance forecast**, each grid point contains a weather element, e.g., 3-hour average precipitation, at the corresponding location ([Japan Meteorological Agency \(2019\)](#)).

Fig. 1(a)subfigure depicts an example of precipitation data drawn on a map, with the vertical and horizontal axes corresponding to the latitude (20° to 50°N) and longitude (120° to 150°E)

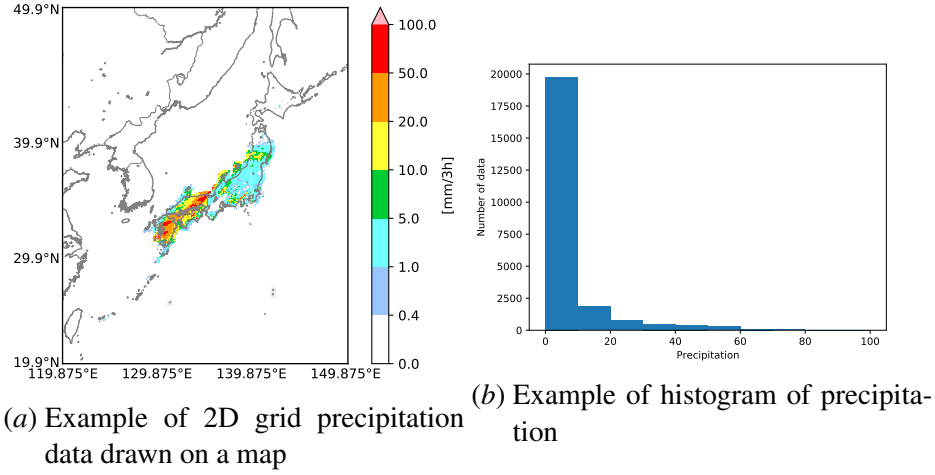


Figure 1: Left: an example of 2D precipitation grid data on July 6th, 15:00UTC, 2018, drawn on a map. In the data, there are 604 and 484 grid points on the vertical and horizontal axes, respectively, and each grid point has a 3-hour average precipitation. Right: the histogram of precipitation data of the left figure. We note that only the data inside the Japanese archipelago are drawn.

respectively. The data are in the form of a 2D grid with each axis sampled at regular intervals of 604 and 484 points. Each grid point contains the average amount of precipitation per three hours.

To improve precipitation forecasts, the methods for integrating multiple guidance forecasts have been actively studied (Mori et al. (2021); Sloughter et al. (2007); Fraley et al. (2010); Liu and Xie (2014); Ji et al. (2019)). In the arithmetic average method (Mori et al. (2021)), three guidance forecasts at each grid point are averaged, where the weight is fixed at  $\frac{1}{3}$  for all grid points. This simple averaging would be expected to reduce the random errors of each forecast.

In the Bayesian average methods (Sloughter et al. (2007); Fraley et al. (2010); Liu and Xie (2014); Ji et al. (2019)), the precipitation is assumed to follow the mixture of distributions, and the weight of each guidance forecast is optimized to minimize the negative log-likelihood approximated with recent past data. Although the weight is adjusted adaptively depending on the seasons of the year in Bayesian averaging methods, a single weight is shared across the 2D grid as well as the arithmetic average. Thus, the local variations of the importance of guidance are not taken into account.

Recently, a deep encoder-decoder image transformation approach called U-Net (Ronneberger et al. (2015)) was applied to directly forecast precipitation on a short-term range, e.g., 30-min to 1-hour, called nowcasting—as a spatial classification task (Agrawal et al. (2020)) and a spatial regression task (Trebing et al. (2020)). These works showed that U-Net based image transformation approaches, e.g., for spatial classification and regression, are promising to directly forecast precipitation in a short term range. However, current U-Net based approaches do not consider the integration of existing multiple forecasts, needed for accurate forecasts for the longer-term range.

For more accurate and long term ranges, we propose U-Net based integration method of multiple guidance forecasts where the input image of multi-channel with guidance forecasts, is transformed

to the weight image, each pixel of which contains a local dependant weight. To this purpose, we extend the architecture of U-Net by introducing

- Truncation loss for stabilizing the training of U-Net, disturbed by the model-misfit errors
- Precipitation Bin Wise (PBW) loss for correcting the imbalance of precipitation data
- Forecast time embedding for adaptively extracting feature maps depending on the target forecast time

Through experiments with the task of the three-hour average precipitation forecast, 3 to 9-hour later, in Japan from April 2018 to March 2019 using the data of JMA ([Japan Meteorological Agency \(a\)](#)), we show the effectiveness of our proposed method in comparison with existing integration methods, individual guidance forecast, and current U-Net-based forecast methods.

## 2. Related works

Let us consider a two-dimensional  $N_{\text{lat}} \times N_{\text{lon}}$  grid of data. Let  $Y_{ij}$ ,  $F_{ijk}$ , and  $W_{ijk}$  be the true value, the  $k$ -th forecast, and the  $k$ -th weight at the grid point  $(i, j)$ .

### 2.1. Integration by arithmetic average

To integrate multiple forecasts with arithmetic averages, all 2D grid forecasts need to have the exact resolution. To this purpose, a down-scaling technique using deep neural networks has been developed by JMA ([Mori et al. \(2021\)](#)). Then, the integrated forecast  $\hat{Y} \in \mathbf{R}^{N_{\text{lat}} \times N_{\text{lon}}}$  is calculated by the weighted sum of forecasts  $F \in \mathbf{R}^{N_{\text{lat}} \times N_{\text{lon}} \times N_g}$  with a fixed weight  $W \in \mathbf{R}^{N_{\text{lat}} \times N_{\text{lon}} \times N_g}$  as follows:

$$\hat{Y} = \sum_{k=1}^{N_g} F_k \otimes W_k \quad (1)$$

$$W_{ijk} = \frac{1}{N_g} \quad \forall ijk \quad (2)$$

where  $\otimes$  is Hadamard product, and  $N_g$  is the number of types of forecasts (see Fig. 2). The improvement of precipitation forecasts was reported thanks to the reduction of random errors in each forecast.

### 2.2. Integration by Bayesian average

With the assumption of the precipitation following the mixture of Bernoulli and Gamma distributions, Bayesian average methods have been actively studied ([Sloughter et al. \(2007\)](#); [Fraleay et al. \(2010\)](#); [Liu and Xie \(2014\)](#); [Ji et al. \(2019\)](#)). Specifically, the binary category of zero precipitation ( $Y_{ij} = 0$ ) or non-zero precipitation ( $Y_{ij} > 0$ ) is assumed to follow the Bernoulli distribution. Besides, the probability that the amount of precipitation ( $Y_{ij} > 0$ ) is below a certain threshold is assumed to follow as a gamma distribution. Then, the probability density function of the precipitation

$Y_{ij}$  is defined using the indicator function  $I(\cdot)$  as

$$p(Y_{ij}|F_{ijk}) = \sum_{k=1}^{N_g} \left[ b(Y_{ij}^{\frac{1}{3}} = 0|F_{ijk}^{\frac{1}{3}})I(Y_{ij} = 0) + \left(1 - b(Y_{ij}^{\frac{1}{3}} = 0|F_{ijk}^{\frac{1}{3}})\right)g_k(Y_{ij}^{\frac{1}{3}}|F_{ijk})I(Y_{ij} > 0) \right] W_k \quad (3)$$

where the cube root is used for reducing the range of the precipitation value.  $b(\cdot)$  is Bernoulli distribution approximated by logistic regression, and  $g_k(\cdot)$  is the gamma distribution of  $k$ -th forecast; the mean  $\mu_{ijk}$  and variance  $\sigma_{ijk}^2$  of which are defined based on the forecast  $F_{ijk}$  as

$$\mu_{ijk} = b_k^0 + b_k^1 F_{ijk}^{\frac{1}{3}}, \quad \sigma_{ijk}^2 = c_k^0 + c_k^1 F_{ijk} \quad (4)$$

where  $\mathbf{b} = (b^0, b^1)$  and  $\mathbf{c} = (c^0, c^1)$  are parameters to be tuned. The integrated forecast  $\hat{Y}$  is calculated as

$$\hat{Y}_{ij} = \text{stat} \left( \sum_{k=1}^{N_g} \left(1 - b(Y_{ij}^{\frac{1}{3}} = 0|F_{ijk}^{\frac{1}{3}})\right)g_k(Y_{ij}^{\frac{1}{3}}|F_{ijk})W_k \right)^3 \quad (5)$$

where  $\text{stat}(\cdot)$  is the statistics of the mixture of distributions, e.g., mean and median. The weight  $W$  is tuned so as to minimize the negative log-likelihood through EM algorithm with the recent past data  $\{Y, F\}$  as

$$\min_{W, \mathbf{c}} - \mathbf{E}_{Y, F} \left[ \log \sum_{i=1}^{N_{\text{lat}}} \sum_{j=1}^{N_{\text{lon}}} p(Y_{ij}|F_{ij}) \right]$$

Various extensions have been made, e.g., for exchangeable and missing forecasts (Fraleay et al. (2010)), and precipitation categories (Ji et al. (2019)).

From this optimization process, the weight could be adjusted adaptively depending on the seasons of the year. However, as well as the arithmetic average, the same weight  $W_k$  is used for the entire grid  $\{i, j\}$  a single weight is shared across the 2D grid for each guidance. Thus, the local variations of the importance of guidance are not taken into account. Besides, various information is available in precipitation forecasts, such as forecast time, current precipitation data, wind speed, and humidity simulated by NWP models. In general, it would not be straightforward to incorporate this additional information in the mixture distribution model due to the nature of linearity.

### 2.3. U-Net based direct approach

In the work (Agrawal et al. (2020)), a sequence of past precipitation images, i.e., observed by 2D-radar, is transformed to a future precipitation image, containing the predicted label of intervals of precipitation at each location. Meanwhile, U-Net could be used as a spatial regression method in the context of precipitation forecast. In the work (Trebing et al. (2020)), U-Net was extended to predict exact precipitation at each location (Trebing et al. (2020)) in 30-min, by introducing attention-mechanisms and depth-wise-separable convolutions to increase the computational efficiency. These works showed that U-Net based image transformation approaches, e.g., for spatial classification and regression, are promising to directly forecast precipitation in a short term range, 30-min to 1-hour.

However, in general, the training of U-Net for a high degree of freedom output, i.e., the regression for the values in an extensive range of precipitation, e.g., 0-100+, and the classification of multiple intervals would be difficult with limited, highly imbalanced, and largely varied data.



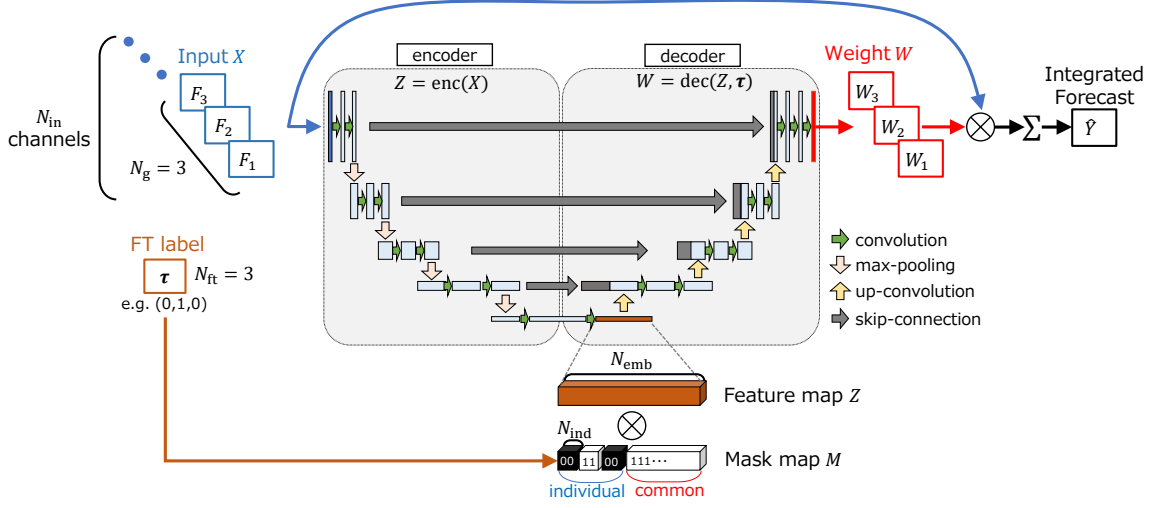


Figure 2: Image to image transformation framework for integrating forecasts using U-Net.

### 3. Proposed integration method

To mitigate these problems, we propose an image transformation based forecast integration method, called *iTraFI*. This method transforms the input image consisting of guidance forecasts and additional information in the channel to the weight image, each pixel of which contains a local dependant weight.

#### 3.1. Architecture of iTraFI

The architecture of our proposed method is illustrated in Fig. 2. Let the input image be  $X \in \mathbf{R}^{N_{\text{lat}} \times N_{\text{lon}} \times N_{\text{in}}}$  which consists of forecast images  $F$  and arbitrary additional information as to its channels. Given the input image  $X$ , the encoder iteratively applies convolution and max-pooling processes to extract the feature map  $Z$ . Given the feature map  $Z$ , the decoder iterates convolution and up-convolution until the resolution of the feature map becomes the same as the one of the input image  $X$ . The weight image  $W$  is outputted through a sigmoidal operation. Then, the weighted sum of the forecast image  $F$  with the weight image  $W$  is used to obtain the forecast image  $\hat{Y}$  as

$$Z = \text{enc}_{\theta}(X), \quad W = \text{dec}_{\phi}(Z, \tau)$$

$$\hat{Y} = \sum_{k=1}^{N_g} F_k \otimes W_k, \quad 0 \leq W_{ijk} \leq 1 \quad (6)$$

where  $\tau$  is the label of forecast time (FT)  $\tau$ , and  $\theta$  and  $\phi$  are parameters of encoder and decoder networks respectively. These parameters are tuned so as to minimize the loss between true value  $Y$  and integrated forecast  $\hat{Y}$  with Frobenius norm regularization as

$$\min_{\theta, \phi} \mathbf{E}_{Y, F} [\text{loss}(Y, \hat{Y}) + \lambda \|W\|_F^2] \quad (7)$$

where  $\lambda$  is the hyper-parameter to adjust the strength of the regularization. The role of Frobenius norm regularization of the weight  $W$  is to reduce the number of combinations of the weight  $W_{ij}$  achieving the minimization of the loss  $\text{loss}(Y_{ij}, \widehat{Y}_{ij})$ . For example, when  $F_{ij} = (30, 40, 100)$  and  $Y_{ij} = 70$ , there are many combinations of  $W_{ij}$  to obtain the zero-error-forecast  $\widehat{Y}_{ij}$ , e.g.,  $W_{ij} = (1.0, 1.0, 0.0)$ ,  $W_{ij} = (1.0, 0.5, 0.2)$ ,  $W_{ij} = (0.0, 0.25, 0.6)$ , and  $W_{ij} = (0.0, 0.0, 0.7)$ . The regularization would lead to the weight  $W_{ij}$  with smaller L2-norm  $\|W_{ij}\|^2$ , e.g.,  $W_{ij} = (0.0, 0.25, 0.6)$ .

### 3.2. Truncation loss

The constraints of  $W_{ijk}$  in Eq. 6 would be useful for visualizing and analyzing the contribution of each forecast  $F_{ijk}$ . However, when the true value  $Y_{ij}$  is the outside of the region spanned by forecasts in training, the optimization problem in Eq. 7 could not be solved. This makes the training of U-Net unstable since the loss at many grid points keeps remained large. On the other hand, removing the constraints on  $W_{ijk}$  would also make the training unstable due to large errors imposed by large disparities between forecasts  $F_{ij}$  and  $Y_{ij}$ . To mitigate these instability problems, we truncate the loss by replacing  $Y_{ij}$  with  $Y_{ij}^*$  as follows:

$$Y_{ij}^* \leftarrow \begin{cases} \max_k F_{ij} & Y_{ij} > \max_k F_{ij} \\ \min_k F_{ij} & Y_{ij} < \min_k F_{ij} \\ Y_{ij} & \text{otherwise} \end{cases} \quad (8)$$

### 3.3. Precipitation Bin Wise Loss

In general, the precipitation data are overly imbalanced. Fig. 1(b)subfigure is the histogram of the precipitation of Fig. 1(a)subfigure. These figures show that while there is medium to heavy rain in the southwest of Japan, zero or near zero precipitation is in most regions. Due to this heavy imbalance, the optimization problem of the parameters  $\theta$  and  $\phi$  (in Eq. 7) with standard loss function, e.g., mean absolute error, is dominated by zero and near-zero precipitation data. Then, this will lead the U-Net not to take the medium to heavy precipitation into account when generating weight  $W$ .

To mitigate this problem, we propose Precipitation Bin Wise (PBW) loss where we split the precipitation into several bins and allocate each data-coordinate  $(i, j)$  into the one of bins according to  $Y_{ij}^*$  and take the average error from the center of bin as

$$\text{loss}(Y^*, \widehat{Y}) \equiv \sum_{t=1}^{N_{\text{thre}}-1} \frac{1}{|D_t|} \sum_{(i,j) \in D_t} \left| \frac{r_t + r_{t+1}}{2} - \widehat{Y}_{ij} \right| \quad (9)$$

$$D_t \equiv \{(i, j) | r_t \leq Y_{ij}^* \leq r_{t+1}\}$$

where  $D_t$  is the set of coordinates  $(i, j)$  and  $r_t$  is the  $t$ -th threshold in the vector  $\mathbf{r}$  defined by

$$\mathbf{r} = (0, 0.5, 1, 3, 5, 10, 15, 20, 25, 30, 40, 50, 60, 70, 80, 100, 120) \quad (10)$$

### 3.4. Forecast time embedding

In order to adaptively generate weight  $W$  depending on the forecast time (FT)  $\tau$ , we propose to select the channels of feature map  $Z$  based on  $\tau$  at the decoder  $\text{dec}(Z, \tau)$ . Specifically, as shown

in Fig. 2, we split the channels of the feature map  $Z \in \mathbf{R}^{w \times h \times N_{\text{emb}}}$  into the individual FT parts ( $N_{\text{ft}}N_{\text{ind}}$  channels) and the common part ( $N_{\text{emb}} - N_{\text{ft}}N_{\text{ind}}$  channels)— $w$  and  $h$ ,  $N_{\text{emb}}$  and  $N_{\text{ind}}$ , denote the width and height of the feature map  $Z$ , the number of channels of  $Z$  and each individual FT part, respectively. The label of FT  $\tau$  is represented by a one-hot vector with the same dimensions as the number of types of FT,  $N_{\text{ft}}$ .

For example, if the FT is  $\{3, 6, 9\}$ , i.e.,  $N_{\text{ft}} = 3$ , then  $\tau$  will be a 3-dimensional vector, and for  $\tau = +6$ ,  $\tau = (0, 1, 0)$ . Then, we create a mask map  $M$  with the same size as the feature map  $Z$ . The first part of the mask map,  $N_{\text{ft}}N_{\text{ind}}$ -channel, is a mask for individual FT parts that the one-hot vector  $\tau$  is extended. The second part,  $N_{\text{emb}} - N_{\text{ft}}N_{\text{ind}}$  channels, is a mask for the common part where all elements are 1. By multiplying this mask map  $M$  by the feature map  $Z$ , we extract the feature map consisting of the FT  $\tau$  and the common parts as

$$Z' = Z \otimes M \quad (11)$$

## 4. Experiment with precipitation forecast

We show the effectiveness of the proposed method, iTraFI, through experiments with the task of three-hour average precipitation forecast, 3 to 9-hour later, from April 2018 to March 2019, in the Japanese archipelago using the data of JMA ([Japan Meteorological Agency \(a\)](#)).

### 4.1. Experimental setting

In terms of computational cost, we resize  $604 \times 484$  of 2D grid data (see Fig. 1(a)subfigure) to  $312 \times 248$  using the resize function of cv2 library. The evaluation period is from April 2018 to March 2019, and the evaluation region is the land and coast of Japan in Fig. 1(a)subfigure. We use three guidance forecasts (LFM, MSM, and GSM) and outputs of NWP models generated at every 3-hour (8 times per day) for the forecast time (FT)  $\{3, 6, 9\}$ —there are  $8 \times 3 = 24$  sets of forecasts per day. We compare the performance of precipitation forecasts generated by the following six methods:

- LFM guidance
- MSM guidance
- GSM guidance
- Arithmetic average (in Sec. 2.1)—integrating the three forecasts using a fixed weight  $W_k = \frac{1}{3}$ .
- Bayesian average (in Sec. 2.2)—training logistic regression model  $b(\cdot)$  and weight  $W_k$  using the recent past 30 days of data  $\{F, Y\}$  based on the study ([Slougher et al. \(2007\)](#)). The parameters of the gamma distribution  $g_k(\cdot)$  in Eq. 4 are tuned manually at  $\mathbf{b} = (0, 1)$  and  $\mathbf{c} = (0, 0.01)$  respectively.
- Proposed method, iTraFI

Table 1: List of channels of input image  $X$  for iTraFI

channel name	source	FT $\tau$	content
LFM	guidance (LFM)	$\{3, 6, 9\}$	Precipitation forecast $F_{k=1}^{(+\tau)}$
MSM	guidance (MSM)	$\{3, 6, 9\}$	Precipitation forecast $F_{k=2}^{(+\tau)}$
GSM	guidance (GSM)	$\{3, 6, 9\}$	Precipitation forecast $F_{k=3}^{(+\tau)}$
OBS	observation	0	Precipitation $Y^{(0)}$
LFM-U	NWP (LFM)	$\{3, 6, 9\}$	EW-component of wind speed at ground [m/s]
LFM-V	NWP (LFM)	$\{3, 6, 9\}$	NS-component of wind speed at ground [m/s]
LFM-RH	NWP (LFM)	$\{3, 6, 9\}$	Humidity at altitude of 925hPa [%]
LFM-PSEA	NWP (LFM)	$\{3, 6, 9\}$	Sea level pressure at ground [hPa]
LFM-OMG	NWP (LFM)	$\{3, 6, 9\}$	Vertical velocity at altitude of 700hPa [hPa/h]
LFM-Z	NWP (LFM)	$\{3, 6, 9\}$	Geopotential height at altitude of 500hPa [m]
LFM-VOR	NWP (LFM)	$\{3, 6, 9\}$	Relative vorticity at altitude of 500hPa [ $10^{-4}/h$ ]

#### 4.2. Setting of iTraFI

In our proposed method, in addition to the guidance forecasts  $F_k^{(+\tau)}$ , arbitrary additional information with the exact resolution could be incorporated as a channel of the input image  $X$ . As shown in Table 1, the current observed precipitation  $Y^{(0)}$  and seven different types of NWP outputs are used as additional information. Therefore, the total number of channels of the input image  $X$  is  $N_{\text{in}} = 10$ —the overall shape is  $312 \times 248 \times 10$ . The number of channels in each layer is set by  $N_{\text{base}} \times 2^{l-1}$  where  $l \in \{1, 2, 3, 4, 5\}$  is the index of layers in each encoder and decoder networks (see Fig. 2), and  $N_{\text{base}} = 20$  is the number of base channels.

For each month of the evaluation period from April 2018 to March 2019, a U-Net model with parameters of  $\text{enc}_\theta(\cdot)$  and  $\text{dec}_\phi(\cdot)$  is trained—12 different models are trained totally. As the training data for a certain target month, we use the data of the same months, and ten days before and after, of the previous three years; the maximum number of epoch for training U-Net is set to 200 and 10% of training data are used for an early stopping operation. The size of the mini-batch is set to 20.

#### 4.3. Evaluation metrics

The Bias score (BI) and Equitable Threat Score (ETS) (Japan Meteorological Agency (2019)) based on the precipitation thresholds as

$$\mathbf{r}' = (0, 0.5, 1, 3, 5, 10, 15, 20, 25, 30, 40, 50, 60, 70, 80, 100) \quad (12)$$

are used to evaluate the performance of precipitation forecasts. The following six sets of grid points are defined based on the threshold  $r_t$ :

$$\begin{aligned} F_{r_t} &\equiv \{(i, j) \mid \widehat{Y}_{ij} \geq r_t\} \\ O_{r_t} &\equiv \{(i, j) \mid Y_{ij} \geq r_t\} \\ FO_{r_t} &\equiv \{(i, j) \mid \widehat{Y}_{ij} \geq r_t \cap Y_{ij} \geq r_t\} \\ FX_{r_t} &\equiv \{(i, j) \mid \widehat{Y}_{ij} \geq r_t \cap Y_{ij} < r_t\} \\ XO_{r_t} &\equiv \{(i, j) \mid \widehat{Y}_{ij} < r_t \cap Y_{ij} \geq r_t\} \\ XX_{r_t} &\equiv \{(i, j) \mid \widehat{Y}_{ij} < r_t \cap Y_{ij} < r_t\} \end{aligned}$$

Since  $FO_{r_t}$  is the set of points where both the forecast  $\widehat{Y}_{ij}$  and the true value  $Y_{ij}$  exceeds the threshold, this set is interpreted as True Positive with respect to the threshold. Similarly, XO, FX and XX correspond to False Negative, False Positive and True Negative respectively.

Using these sets, BI and ETS are defined as

$$BI_{r_t} \equiv \frac{|FO_{r_t}| + |FX_{r_t}|}{|FO_{r_t}| + |XO_{r_t}|} \quad (13)$$

$$ETS_{r_t} \equiv \frac{|FO_{r_t}| - S_{r_t}}{|FO_{r_t}| + |FX_{r_t}| + |XO_{r_t}| - S_{r_t}}, \quad S_{r_t} \equiv \frac{|O_{r_t}|}{N} |F_{r_t}| \quad (14)$$

where  $N$  is the number of all grid points, and  $S_{r_t}$  represents the number of FO in the random forecast. In other words, BI is the ratio of a forecast concerning the observation, and ETS is the ratio of true positive over the forecast and observation, discounted by the random forecast.

#### 4.4. Evaluation results

Fig. 3 and Fig. 4 depict the performance of precipitation forecast with different settings of FTs  $\tau$  using BI and ETS—for visualization purposes, ETS values are divided by ones of arithmetic average.

Fig. 3(b)subfigure shows that the existing methods, arithmetic and Bayesian average methods, tend to provide better performances than each guidance forecast, i.e., LFM, MSM, or GSM, in almost all precipitation thresholds except for the LFM with the threshold of 100 [mm/3h]. This indicates that random error in each of guidance forecasts, could be canceled to some extent by averaging three forecast values.

However, large errors in one of three forecasts or correlated errors among two or three forecasts could not be reduced in principle, especially with the constant weight  $W_k$ . Fig. 5 depicts forecasts generated by LFM-, MSM-, GSM-guidance, arithmetic average, Bayesian average, and proposed method on July 6th, 9:00UTC, 2018 for FT  $\tau = 6$  where the true precipitation value is depicted in Fig. 1(a)subfigure. Guidance forecasts by LFM, MSM, and GSM in Fig. 5(a)subfigure, 5(b)subfigure, and 5(c)subfigure seem to forecast well when comparing with Fig. 1(a)subfigure but there are errors in some details. For example, there exist large XO (False Negative) regions in Hyogo (around 34.5°N & 135°E) and FX (False Positive) regions in Chugoku (around 34.5°N & 133°E) in GSM-guidance.

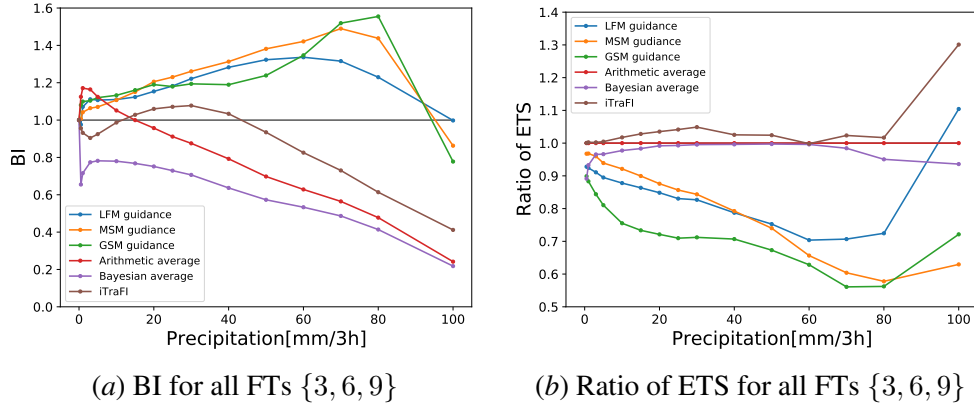


Figure 3: Performance of precipitation forecast with all forecast times (FTs)  $\tau \in \{3, 6, 9\}$  for LFM-, MSM-, GSM-guidance, arithmetic average, Bayesian average and proposed method (iTraFI). Forecasts are done every 3 hours from April 2018 to March 2019. As the metric of precipitation forecast, the BI and ETS (see Sec. 4.3) are calculated at each precipitation threshold  $r'$  (in Eq. 12). For the visualization purpose, each ETS values are divided by the ones of arithmetic average.

Forecasts obtained by integrating three forecasts using arithmetic average in Fig. 5(d)subfigure or Bayesian average in Fig. 5(e)subfigure show that some errors in guidance forecasts could be mitigated where the FX regions of the GSM are improved in both arithmetic and Bayesian average methods. However, the significant errors in guidance forecasts keep remaining, i.e., the XO regions of the GSM in both methods. These examples indicate that arithmetic and Bayesian average methods would have problems of the trade-off between regions in which the error in one region improves but it gets worse in other regions since a single weight  $W_k$  is shared over the entire region.

On the other hand, our proposed method, iTraFI, would enable us to adaptively tune the weight depending on the location, and Fig. 3 and Fig. 4 show that it further improves the performance more than existing methods. That is, the BI values of the proposed method are more close to one as shown in Fig. 3(a)subfigure, 4(a)subfigure, 4(c)subfigure, and 4(e)subfigure, indicating that the proposed method forecasts precipitation at each threshold with a moderate number of times. The ETS values of the proposed method are higher in almost all thresholds as shown in Fig. 3(b)subfigure, 4(b)subfigure, 4(d)subfigure, and 4(f)subfigure, indicating that the proposed method provides more accurate integrated forecasts than existing methods. Fig. 5(h)subfigure and Fig. 6 depict the integrated forecasts and the weights, obtained by our proposed method. These figures show that the XO and FX regions of the GSM are removed in the integrated forecasts (in Fig. 5(h)subfigure) since the weights in the regions are set at low values, i.e., almost zero, as shown in Fig. 6(c)subfigure. Meanwhile, the FO (True Positive) forecasts of those regions in the LFM are accepted since the weights of the regions in LFM are set high values, i.e., about 0.6 as shown in Fig. 5(a)subfigure and Fig. 6(a)subfigure.

Fig. 6(d)subfigure depicts the normalized histogram of the highest weights among three guidance forecasts at each location, in the case that the FT  $\tau = 6$  true precipitation  $\leq 69$  and  $\geq 70$  in July 2018. The figure shows that our proposed method could set weights adaptively i.e., higher/lower weights for higher/lower future precipitation.

#### 4.5. Ablation study

To confirm the effects of the architecture and components of our proposed method, iTraFI, we compare the performance of iTraFI with the following variants;

- iTraFI without additional information (see Table. 1)—only guidance forecasts,  $F$ , are used as the input.
- iTraFI without PBW (see Sec. 3.3)—the mean absolute error (MAE) is used as the loss function instead of PBW loss.
- iTraFI without truncation loss (see Sec. 3.2)—true values  $Y$  are directly used for computing PBW loss in the training.
- iTraFI without forecast time embedding (see Sec. 3.4)—the feature map  $Z$  is passed to the decoder as it is.
- U-Net based regression (Trebing et al. (2020))—the precipitation value at each location is directly estimated using the same U-Net architecture and input as iTraFI, by minimizing the MAE loss function.
- U-Net based classification (Agrawal et al. (2020))—a precipitation interval consisting of adjacent thresholds in  $\mathbf{r}'$  (Eq. 12), is selected, i.e., 16-class classification task, using the same U-Net architecture and input as iTraFI, by minimizing the cross-entropy loss function. To convert to the precipitation value, the lower bound of the selected interval is used.

Fig. 7(a)subfigure and Fig. 7(b)subfigure depicts BI and ETS values of iTraFI and its variants. The figures show that removing the PBW loss or truncation loss of iTraFI largely decreases its performance, although removing additional information and forecast time embedding would slightly affect the performance. This indicates that the PBW loss could treat well prohibitive imbalance problems that inevitably happened in the precipitation data, and the truncation loss would contribute to stabilizing the training of U-Net. Fig. 7(c)subfigure and Fig. 7(d)subfigure depict BI and ETS values of iTraFI and iTraFI w/o forecast time embedding in each of FTs. The figures show that forecast time embedding would be effective except for some cases; high precipitation in FT  $\tau = 6$  and 9.

In addition, Fig. 7(b)subfigure shows that the U-Net based regression and classification methods decrease the performance of forecasts. In general, the training of deep neural networks with a high degree of freedom output, i.e., the regression for the values in a range of 0-100+ precipitation and the classification of 16 intervals would be complicated with limited, highly imbalanced, and largely varied data. The task of precipitation forecast would be a typical example of such cases. Fig. 5(f)subfigure and Fig. 5(g)subfigure show that although the tendency of predicted precipitation is similar to the true precipitation in Fig. 1(a)subfigure, the predicted values tend to be lower and thus there would be lots of XO (False Negative) regions, indicating a naive application of existing U-Net would not help for accurate precipitation forecasts, especially for a longer-term range.

Overall, these experimental results show that our proposed architecture of iTraFI generating weights for integrating guidance forecasts could be effective as our proposed method outperforms existing methods.



## 5. Conclusion

In this work, we propose a new approach, called iTraFI for integrating precipitation forecasts, by converting forecasts images to weight images; this enables us to tune the weight adaptively depending on the locations and to incorporate additional information from numerical weather prediction models and forecast time. Through experiments with the task of precipitation forecast for the future time of 3, 6, and 9 in Japan, we show the effectiveness of our proposed method over existing methods, including arithmetic average, Bayesian average, and recent U-Net-based regression and classification methods. The further improvement for a longer forecast time than 9-hour would be expected as future works. In addition, our proposed method is not specific to the precipitation data and could be applicable to other forecast tasks with multiple spatial forecasts, e.g., wind speed and temperature, etc. The evaluation in other forecast tasks would also be future works.

## References

- S. Agrawal, L. Barrington, C. Bromberg, J. Burge, C. Gazen, and J. Hickey. Machine learning for precipitation nowcasting from radar images. 2020.
- Chris Fraley, Adrian E. Raftery, and Tilmann Gneiting. Calibrating multimodel forecast ensembles with exchangeable and missing members using bayesian model averaging, 2010.
- Japan Meteorological Agency. Japan meteorological business support center. <http://www.jmbssc.or.jp/en/index-e.html>, a.
- Japan Meteorological Agency. Annual precipitation in japan. [https://www.data.jma.go.jp/cpdinfo/temp/an\\_jpn\\_r.html](https://www.data.jma.go.jp/cpdinfo/temp/an_jpn_r.html), b.
- Japan Meteorological Agency. *Outline of the operational numerical weather prediction at the Japan Meteorological Agency*. Japan Meteorological Agency, Ministry of Land, Infrastructure, Transport and Tourism, 2019.
- L. Ji, X. Zhi, S. Zhu, and K. Fraedrich. Probabilistic precipitation forecasting over east asia using bayesian model averaging, 2019.
- J. Liu and Z. Xie. Bma probabilistic quantitative precipitation forecasting over the huaihe basin using tidge multimodel ensemble forecasts, 2014.
- Y. Mori, N. Fujikane, and T. Inoue. Ongoing development of methods to combine forecast guidance outputs using artificial intelligence technology at the japan meteorological agency. In *AMS Annual Meeting*, 2021.
- O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer Assisted Intervention (MICCAI2015)*, pages 234–241, 2015.
- J. M. Sloughter, A. E. Raftery, T. Gneiting, and C. Fraley. Probabilistic quantitative precipitation forecasting using bayesian model averaging, 2007.
- K. Trebing, T. Stanczyk, and S. Mehrkanon. Smaat-unet: Precipitation nowcasting using a small attention-unet architecture. 2020.

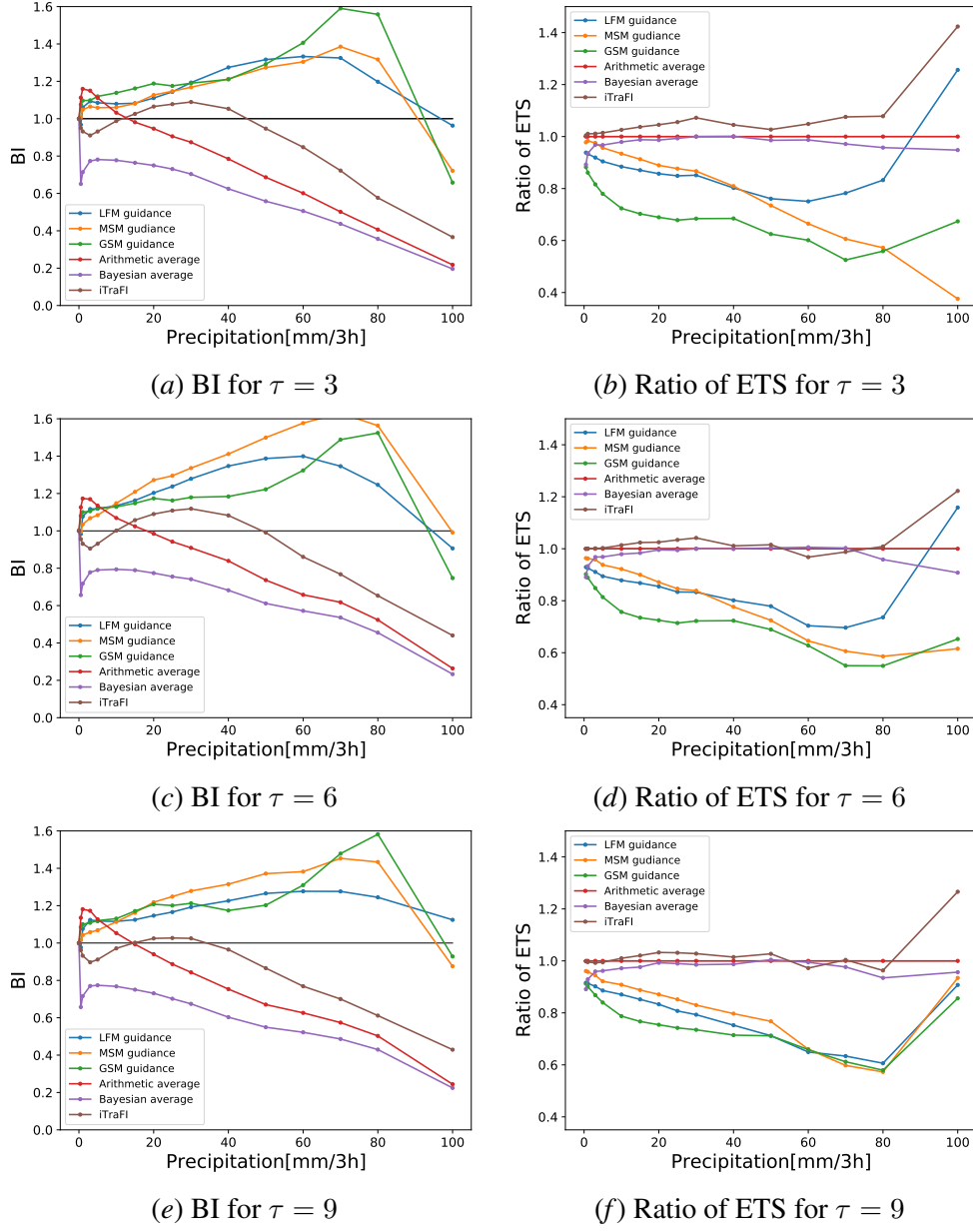


Figure 4: Performance of precipitation forecast with each individual forecast time (FT)  $\tau \in \{3, 6, 9\}$  for LFM-, MSM-, GSM-guidance, arithmetic average, Bayesian average and our proposed method, iTraFI. Forecasts are done every 3 hours from April, 2018 to March, 2019. As the metric of precipitation forecast, the BI and ETS (see Sec. 4.3) are calculated at each precipitation threshold  $r'$  (in Eq. 12). For the visualization purpose, ETS values are divided by the ones of arithmetic average.

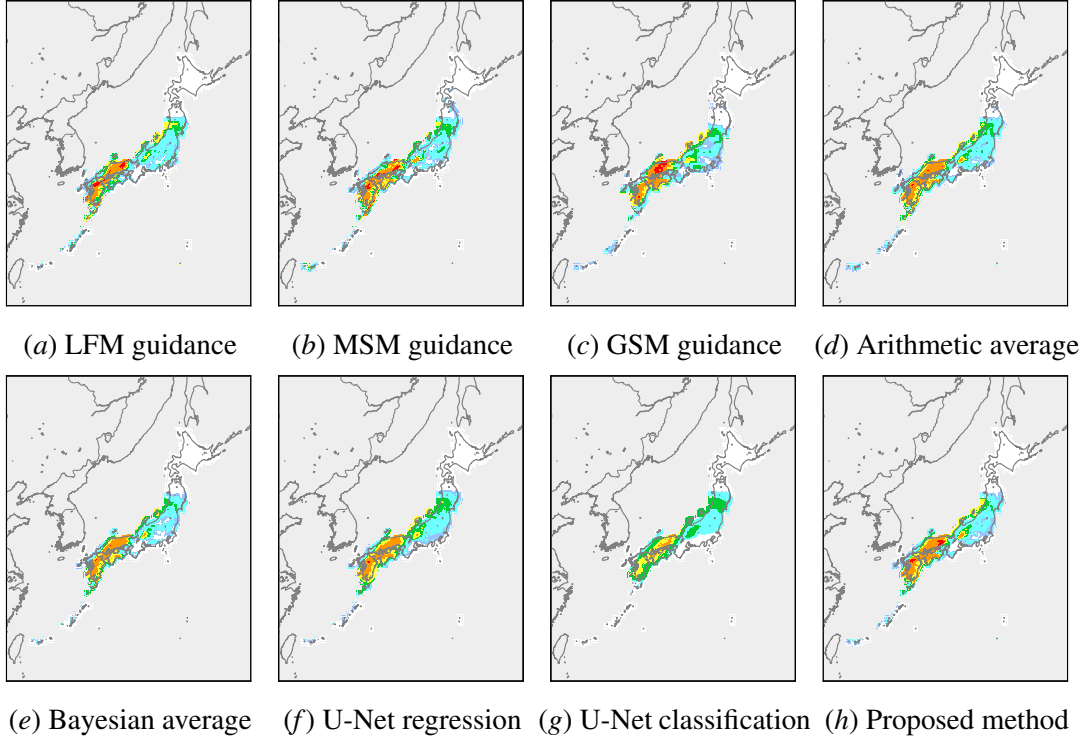


Figure 5: Precipitation forecasts drawn on a map. The forecasts are generated by LFM-, MSM-, and GSM-guidance, arithmetic average, Bayesian average, U-Net based regression (Trebing et al. (2020)) and classification (Agrawal et al. (2020)) (see Sec. 4.5), and proposed method (iTraFI) on July 6th, 9:00UTC, 2018 for FT  $\tau = 6$ . The true precipitation data are shown in Fig. 1(a)subfigure.

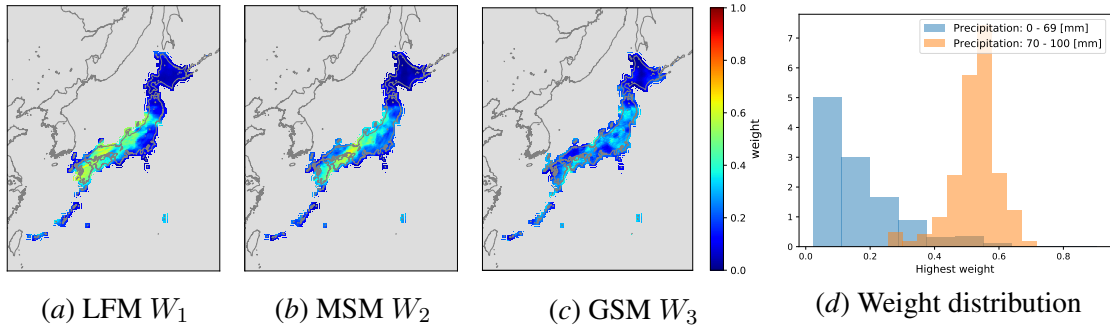


Figure 6: Left: Weights  $W_k$  for each guidance, drawn on a map. These weights are generated by proposed method, iTraFI, on July 6th, 9:00UTC, 2018 for the FT  $\tau = 6$ . Right: histogram of highest weights among three guidance forecasts at each location for the FT  $\tau = 6$  true precipitation  $\leq 69$  and  $\geq 70$ .

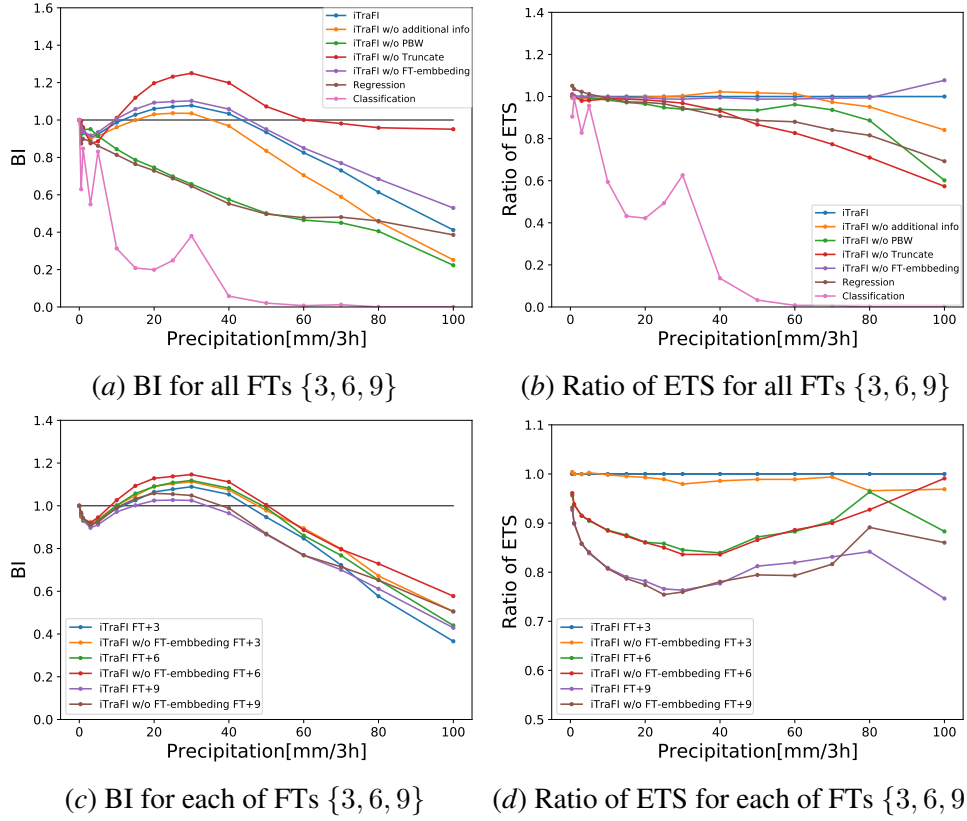


Figure 7: Ablation study—Fig. 7(a)subfigure and Fig. 7(b)subfigure show the performance comparison of our proposed method, iTraFI, with w/o additional information, w/o PBW, w/o truncation loss, w/o forecast time embedding, U-Net based precipitation regression (Trebing et al. (2020)), and precipitation interval classification (Agrawal et al. (2020)). Fig. 7(c)subfigure and Fig. 7(d)subfigure show the performance comparison of iTraFI and iTraFI w/o forecast time embedding in each of FTs.