

POSTRAINBENCH: A COMPREHENSIVE BENCHMARK AND A NEW MODEL FOR PRECIPITATION FORECASTING

Yujin Tang¹ Jiaming Zhou¹ Xiang Pan² Zeying Gong¹ Junwei Liang^{1,3} *

¹AI Thrust, The Hong Kong University of Science and Technology (Guangzhou)

²School of Atmospheric Science, Nanjing University

³Department of Computer Science and Engineering, The Hong Kong University of Science and Technology
tangyujin0275@gmail.com, junweiliang@hkust-gz.edu.cn

ABSTRACT

Accurate precipitation forecasting is a vital challenge of societal importance. Though data-driven approaches have emerged as a widely used solution, solely relying on data-driven approaches has limitations in modeling the underlying physics, making accurate predictions difficult. We focus on the Numerical Weather Prediction (NWP) post-processing based precipitation forecasting task to couple Machine Learning techniques with traditional NWP. This task remains challenging due to the imbalanced precipitation data and complex relationships between multiple meteorological variables. To address these limitations, we introduce the **PostRainBench**, a comprehensive multi-variable NWP post-processing benchmark, and **CAMT**, a simple yet effective Channel Attention Enhanced Multi-task Learning framework with a specially designed weighted loss function. Extensive experimental results on the proposed benchmark show that our method outperforms state-of-the-art methods by 6.3%, 4.7%, and 26.8% in rain CSI and improvements of 15.6%, 17.4%, and 31.8% over NWP predictions in heavy rain CSI on respective datasets. Most notably, our model is the first deep learning-based method to outperform NWP approaches in heavy rain conditions. These results highlight the potential impact of our model in reducing the severe consequences of extreme rainfall events. Our datasets and code are available at <https://github.com/yuyujintang/PostRainBench>.

1 INTRODUCTION

Precipitation forecasting (Sønderby et al., 2020; Espeholt et al., 2022) refers to the problem of providing a forecast of the rainfall intensity based on radar echo maps, rain gauge, and other observation data as well as the Numerical Weather Prediction (NWP) models (Shi et al., 2017). Accurate rainfall forecasts can guide people to make optimal decisions in production and life. Though the occurrence of extreme precipitation events is relatively infrequent, they can lead to adverse impacts on both agricultural production and community well-being (de Witt et al., 2021).

In the past few years, geoscience has begun to use deep learning to better exploit spatial and temporal structures in the data. Comparing to directly extrapolating rainfall field with convolutional–recurrent approaches (Shi et al., 2015; Wang et al., 2017; Shi et al., 2017), which is mainly based on data-driven extrapolation and lacks physics-based modeling (Kim et al., 2022), post-processing NWP rainfall prediction provides more physically-consistent results. Combining AI-based and NWP methods can bring about both strengths for a stronger performance (Bi et al., 2023). For the post-processing task, NWP predictions are fed to a deep learning model which is trained to output refined precipitation forecasts, while rainfall station observations are used as ground truth. In a nutshell, the overall task is to post-process the predictions from NWP with deep models, under the supervision of rainfall observations.

However, Post-NWP optimization poses several distinct challenges that distinguish it from typical weather forecasting optimization problems and computer vision tasks. (1) The absence of a unified

*Corresponding author

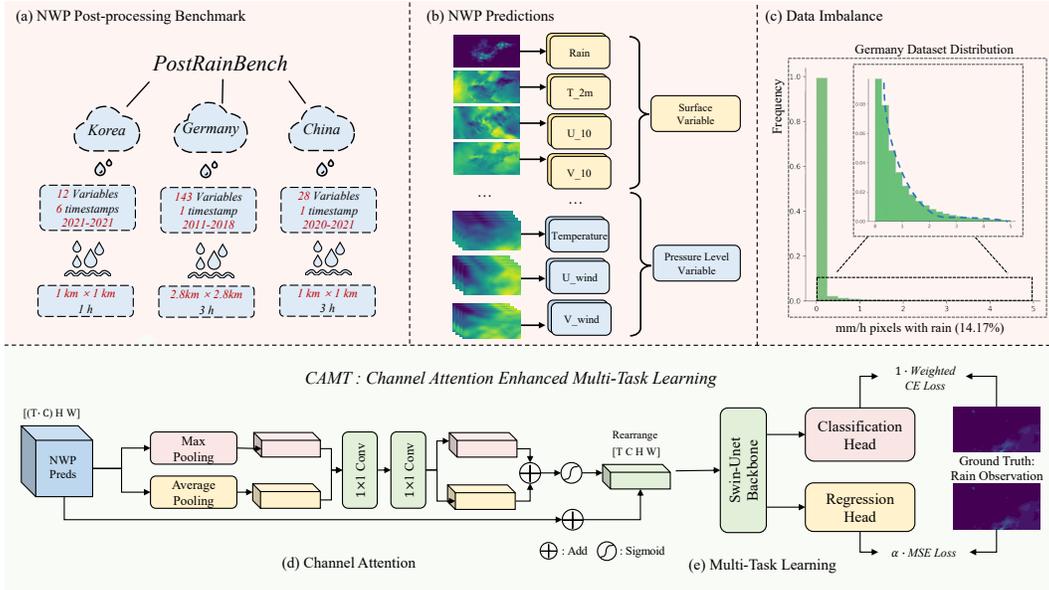


Figure 1: An overview of the proposed **PostRainBench** and **CAMT** framework. (a) benchmark’s attributes. (b) input composition. (c) distribution of the German dataset. The bottom section illustrates our CAMT workflow: (d) NWP inputs undergo processing by the Channel Attention Module, followed by a Swin-Unet backbone. (e) Multi-task learning with hybrid weighted loss.

benchmark hindering cross-model evaluation; (2) The uncertainty of variable selection and modeling arises from the spatial dependencies and diverse statistical properties of atmospheric variables; (3) The existing significant data imbalance, i.e., light rain and heavy rain, makes the task harder to implement.

To tackle the aforementioned challenges, we introduce **PostRainBench**, a comprehensive multi-variable benchmark, which covers the full spectrum of scenarios with and without temporal information and various combinations of NWP input variables and we propose a new model learning framework **CAMT**, a simple yet effective Channel Attention Enhanced Multi-task Learning framework with a specially designed weighted loss function. On the proposed benchmark, our model outperforms state-of-the-art methods in rain Critical Success Index(CSI) on three datasets. Furthermore, it’s worth highlighting a significant milestone achieved by our model. In heavy rain, with improvements of **15.6%**, **17.4%**, and **31.8%** in CSI over NWP predictions across respective datasets. This underscores its potential to effectively mitigate substantial losses in the face of extreme weather events.

2 DATA AND METHODOLOGY

2.1 DATASETS

Our benchmark **PostRainBench** is comprised of three datasets, two of which are sourced from prior research, while the third is collected from a public challenge. The first dataset, called KoMet (Kim et al., 2022), was collected in South Korea. The input data originates from GDAPS-KIM, a global numerical weather prediction model that furnishes hourly forecasts for diverse atmospheric variables. The second dataset originates from Germany (Rojas-Campos et al., 2022). The input data is derived from the COSMO-DE-EPS forecast (Peralta et al., 2012), which provides 143 variables of the atmospheric state. The third dataset originates from China and provides hourly, 1 km × 1 km resolution, 3-hour grid point precipitation data for the rainy season. It includes 3-hour lead time forecasts from a regional NWP model, with 28 surface and pressure level variables. We summarize important details of the three datasets in the Table 3 and analyze the distribution of the observed precipitation data in Table 5. All three datasets exhibit significant imbalances, which presents a great challenge to predict extreme weather scenarios.

2.2 TASK DEFINITION

In this study, we consider optimizing the following model:

$$\min_{\mathbf{w}} \left\{ \mathcal{L}(\mathbf{w}; \mathcal{D}) \triangleq \mathbb{E}_{(X_t, y_t) \sim \mathcal{D}} [\ell(y_t; F(X_t, \mathbf{w}))] \right\} \quad (1)$$

where \mathcal{L} represents the objective function parameterized by \mathbf{w} on the dataset \mathcal{D} . The input is NWP predictions X_t , the corresponding ground-truth is rain observation y_t at time t , and ℓ denotes the loss function between the output of our proposed model $F(\cdot, \mathbf{w})$ and the ground-truth.

2.3 METHOD

As illustrated in Figure 1, our model can be divided into three parts. The first part is a channel attention module (Woo et al., 2018). The second part is the Swin-Unet backbone (Cao et al., 2022a) that generates linear projections. The third part is a multi-task learning branch with a hybrid loss. We describe the first and third parts in detail below and put explanations of Swin-Unet in Section A.3.1 and A.3.2.

We introduce the Channel Attention Module (CAM), which enables variable selection for a unified NWP post-processing task, and models intricate relationships between variables. CAM aggregates spatial information of a feature map by using both average-pooling and max-pooling operations, generating two different spatial context descriptors: $\mathbf{F}_{\text{avg}}^c$ and $\mathbf{F}_{\text{max}}^c$. Both descriptors are forwarded to a shared multi-layer perceptron (MLP) to produce a channel attention map $\mathbf{M}_c \in \mathbb{R}^{C \times 1 \times 1}$.

For model optimization, we introduce a combination of Mean Squared Error (MSE) loss and weighted Cross-Entropy (CE) loss within a multi-task learning framework, incorporating two task outputs $\tilde{\mathbf{y}}_{cls}$, $\tilde{\mathbf{y}}_{reg}$ with a hyperparameter α .

Utilizing dedicated classification and regression heads encourages the backbone to focus on learning essential features for both tasks. As previously mentioned, precipitation forecasting grapples with the challenge of highly imbalanced class distributions from a classification standpoint. To tackle this issue, we apply class weights w_c based on the class distribution of each dataset. The full loss function L_{hybrid} is defined as:

$$L_{hybrid} = L_{cls} + \alpha L_{reg}, \quad L_{cls} = \sum_{i=1}^h \sum_{j=1}^w \left(- \sum_{c=1}^M w_c y_t \log(\tilde{\mathbf{y}}_{cls}) \right), \quad L_{reg} = \sum_{i=1}^h \sum_{j=1}^w (\tilde{\mathbf{y}}_{reg} - y_t)^2 \quad (2)$$

3 EXPERIMENTS

We compare our proposed Swin-Unet-based CAMT framework with various strong baselines, including the NWP method, three deep learning models (ConvLSTM, UNet, MetNet). Swin-Unet (Ronneberger et al., 2015) is a Unet-like Transformer. The tokenized image patches are fed into the Swin Transformer-based (Liu et al., 2021) U-shaped Encoder-Decoder architecture with skip connections for local-global semantic feature learning.

3.1 RESULTS

As shown in Table 1, for the Korea dataset, our method demonstrates an improvement of 6.3% in rain prediction CSI compared to the state-of-the-art (SOTA) approach, which is ConvLSTM. We highlight that CAMT achieves a remarkable 15.6% improvement in heavy rain prediction CSI over the NWP method, which is the first DL model to surpass NWP results for extreme weather conditions.

For the Germany dataset, U-Net emerges as the top performer among previous models, particularly excelling in rain CSI. Notably, our method achieves a 4.7% improvement over U-Net. When it comes to heavy rain prediction, U-Net’s performance is limited and the NWP model outperforms

Table 1: Experimental Results on the proposed PostRainBench. Each model undergoes three runs with different random seeds, and we report the mean, standard deviation (std), and best performance in terms of CSI and Heidke Skill Score(HSS). The best results are highlighted in **bold**, with the second-best results underlined. We report the relative improvement of our method (Swin-Unet+CAMT) over the best result among the baselines and NWP. In the context of the results, '↑' indicates that higher scores are better.

		Rain				Heavy Rain			
		CSI↑		HSS↑		CSI↑		HSS↑	
		Mean(Std)	Best	Mean(Std)	Best	Mean(Std)	Best	Mean(Std)	Best
Korea	NWP	0.263(±0.000)		*		0.045(±0.000)		*	
	U-Net	0.300(±0.025)	<u>0.322</u>	0.384 (±0.025)	0.408	0.006(±0.005)	0.010	0.011(±0.009)	0.018
	ConvLSTM	<u>0.302</u> (±0.009)	0.312	0.384 (±0.009)	<u>0.395</u>	0.009(±0.007)	<u>0.015</u>	<u>0.016</u> (±0.012)	<u>0.026</u>
	MetNet	0.298(±0.012)	0.307	0.375(±0.014)	0.384	0.005(±0.007)	0.012	0.009(±0.012)	0.023
	Ours	0.321 (±0.005)	0.326	0.384 (±0.007)	0.389	0.052 (±0.010)	0.058	0.089 (±0.017)	0.097
	Ours Δ	+6.3%		+0%		+15.6%		+456.3%	
Germany	NWP	0.338(±0.000)		0.252(±0.000)		<u>0.178</u> (±0.000)		<u>0.173</u> (±0.000)	
	U-Net	<u>0.491</u> (±0.007)	<u>0.495</u>	<u>0.601</u> (±0.006)	<u>0.605</u>	0.082(±0.028)	0.107	0.148(±0.048)	0.189
	ConvLSTM	0.477(±0.026)	0.478	0.587(±0.004)	0.590	0.091(±0.041)	<u>0.121</u>	0.162(±0.068)	<u>0.212</u>
	MetNet	0.485(±0.002)	0.487	0.595(±0.005)	0.599	0.027(±0.016)	0.094	0.147(±0.027)	0.168
	Ours	0.514 (±0.003)	0.518	0.609 (±0.006)	0.616	0.209 (±0.014)	0.224	0.339 (±0.020)	0.359
	Ours Δ	+4.7%		+1.3%		+17.4%		+96.0%	
China	NWP	<u>0.164</u> (±0.000)		<u>0.123</u> (±0.000)		<u>0.110</u> (±0.000)		0.089(±0.000)	
	U-Net	0.065(±0.007)	0.073	0.093(±0.009)	0.103	0.058(±0.014)	0.070	0.089(±0.024)	0.110
	ConvLSTM	0.054(±0.011)	0.066	0.079(±0.009)	0.088	0.065(±0.003)	0.068	<u>0.104</u> (±0.010)	0.114
	MetNet	0.064(±0.019)	<u>0.078</u>	0.061(±0.047)	<u>0.106</u>	0.057(±0.017)	<u>0.076</u>	0.069(±0.057)	<u>0.118</u>
	Ours	0.208 (±0.007)	0.216	0.274 (±0.014)	0.289	0.145 (±0.015)	0.163	0.225 (±0.019)	0.246
	Ours Δ	+26.8%		+122.8%		+31.8%		+116.3%	

* For Korea dataset, NWP method's HSS is not reported. For all NWP method, we only have the mean value.

all previous DL models. Our method shows a substantial 17.4% improvement over NWP, marking a significant advancement.

In the case of the China dataset, the NWP method demonstrates better performance in both rain and heavy rain prediction compared to previous DL models. Our method achieves improvements of 26.8% and 31.8% over the NWP method under these two conditions, respectively.

3.2 ABLATION STUDY

We conduct an ablation study by systematically disabling certain components of our CAMT Component and evaluating the CSI results for both rain and heavy rain in Table 2. Specifically, we focus on the weighted loss, multi-task learning, and channel attention modules as these are unique additions to the Swin-Unet backbone. In the first part, we use Swin-Unet with CAMT framework (a) as a baseline and we disable each component in CAMT and demonstrate their respective outcomes. In the second part, we use Swin-Unet without CAMT framework (e) as a baseline and we gradually add each component to the model to understand its role.

Weighted Loss (b) Without the weighted Loss in CAMT, there is a slight increase in rain CSI, but heavy rain CSI shows a dominant 97.6% decrease. (f) Adding the weighted loss to Swin-Unet results in a 6.0% decrease in rain CSI, but a significant improvement in heavy rain CSI.

Multi-Task Learning (c) Without multi-task learning, there is a 3.7% drop in rain CSI, along with a notable 8.1% decrease in heavy rain CSI. (g) Incorporating multi-task learning into Swin-Unet leads to a comparable performance of rain CSI but brings a slight increase in heavy rain CSI.

CAM (d) In the absence of CAM, we observe a 1.8% decrease in rain CSI and a significant 11.1% decrease in heavy rain CSI. (h) The introduction of CAM into Swin-Unet leads to a rain CSI similar

to the baseline but demonstrates an impressive 11.5% improvement in heavy rain CSI. It indicates that CAM is effective for selecting and modeling multiple weather variables.

Table 2: Ablation study on Germany dataset (Rojas-Campos et al., 2022). We disable components of the framework in each experiment and report rain and heavy rain CSI as the evaluation metric.

	Weighted Loss	Multi-Task Learning	CAM	Rain		Heavy Rain	
				CSI↑	HSS↑	CSI↑	HSS↑
(a)	✓	✓	✓	0.514	0.609	0.209	0.339
(b)	✗	✓	✓	0.517 (+0.6%)	0.625 (+2.6%)	0.042 (−97.6%)	0.008 (−11.1%)
(c)	✓	✗	✓	0.495 (−3.7%)	0.588 (−3.4%)	0.192 (−8.1%)	0.317 (−6.5%)
(d)	✓	✓	✗	0.505 (−1.8%)	0.602 (−1.1%)	0.183 (−11.1%)	0.305 (−11.1%)
(e)	✗	✗	✗	0.521	0.628	0.000	0.000
(f)	✓	✗	✗	0.490 (−6.0%)	0.580 (−7.6%)	0.188 ↑↑↑	0.307 ↑↑↑
(g)	✗	✓	✗	0.516 (−0.1%)	0.629 (+0.2%)	0.067 ↑	0.007 ↑
(h)	✗	✗	✓	0.513 (−1.5%)	0.624 (−0.6%)	0.115 ↑↑	0.204 ↑↑

Although Swin-UNet can achieve a relatively high CSI when used alone (e), it does not have the ability to predict heavy rain. Importantly, these three enhancements complement each other. Weighted loss and multi-task learning are effective in improving simultaneous forecasting under the unbalanced distribution of light rain and heavy rain, while CAM provides comprehensive improvements.

4 CONCLUSION

In this paper, we introduce **PostRainBench**, a comprehensive multi-variable benchmark for NWP post-processing-based precipitation forecasting and we present **CAMT**, Channel Attention Enhanced Multi-task Learning framework with a specially designed weighted loss function. Our approach demonstrates outstanding performance improvements compared to the three baseline models and the NWP method. In conclusion, our research provides novel insights into the challenging domain of highly imbalanced precipitation forecasting tasks. We believe our benchmark could help advance the model development of the research community.

5 ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (No. 62306257). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the National Natural Science Foundation.

Reproducibility Statement Our code and datasets are available at <https://github.com/yyujintang/PostRainBench>.

REFERENCES

- Lindsey R Barnes, David M Schultz, Eve C Grunfest, Mary H Hayden, and Charles C Benight. Corrigendum: False alarm rate or false alarm ratio? *Weather and Forecasting*, 24(5):1452–1454, 2009.
- Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, pp. 1–6, 2023.
- Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pp. 205–218. Springer, 2022a.
- Yuan Cao, Lei Chen, Danchen Zhang, Leiming Ma, and Hongming Shan. Hybrid weighting loss for precipitation nowcasting from radar images. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3738–3742. IEEE, 2022b.
- Christian Schroeder de Witt, Catherine Tong, Valentina Zantedeschi, Daniele De Martini, Alfredo Kalaitzis, Matthew Chantry, Duncan Watson-Parris, and Piotr Bilinski. Rainbench: Towards data-driven global precipitation forecasting from satellite imagery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 14902–14910, 2021.
- RJ Donaldson, Rosemary M Dyer, and Michael J Kraus. Objective evaluator of techniques for predicting severe weather events. In *Bulletin of the American Meteorological Society*, volume 56, pp. 755–755. AMER METEOROLOGICAL SOC 45 BEACON ST, BOSTON, MA 02108-3693, 1975.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- Lasse Espeholt, Shreya Agrawal, Casper Sønderby, Manoj Kumar, Jonathan Heek, Carla Bromberg, Cenk Gizen, Rob Carver, Marcin Andrychowicz, Jason Hickey, et al. Deep learning for twelve hour precipitation forecasts. *Nature communications*, 13(1):1–10, 2022.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
- Robin J Hogan, Christopher AT Ferro, Ian T Jolliffe, and David B Stephenson. Equitability revisited: Why the “equitable threat score” is not equitable. *Weather and Forecasting*, 25(2):710–726, 2010.
- Taehyeon Kim, Namgyu Ho, Donggyu Kim, and Se-Young Yun. Benchmark dataset for precipitation forecasting by post-processing the numerical weather prediction. *arXiv preprint arXiv:2206.15241*, 2022.
- Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.
- Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, et al. Four-castnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*, 2022.
- C Peralta, Z Ben Bouallègue, SE Theis, C Gebhardt, and M Buchhold. Accounting for initial condition uncertainties in cosmo-de-eps. *Journal of Geophysical Research: Atmospheres*, 117 (D7), 2012.

- Markus Reichstein, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalho, and fnm Prabhat. Deep learning and process understanding for data-driven earth system science. *Nature*, 566(7743):195–204, 2019.
- Adrian Rojas-Campos, Michael Langguth, Martin Wittenbrink, and Gordon Pipa. Deep learning models for generation of precipitation maps based on numerical weather prediction. *EGUsphere*, pp. 1–20, 2022.
- Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pp. 234–241. Springer, 2015.
- U Schättler, G Doms, and C Schraff. A description of the nonhydrostatic regional cosmo-model part vii: user’s guide. *Deutscher Wetterdienst Rep. COSMO-Model*, 4:142, 2008.
- Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015.
- Xingjian Shi, Zhihan Gao, Leonard Lausen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun Woo. Deep learning for precipitation nowcasting: A benchmark and a new model. *Advances in neural information processing systems*, 30, 2017.
- Casper Kaae Sønderby, Lasse Espeholt, Jonathan Heek, Mostafa Dehghani, Avital Oliver, Tim Salimans, Shreya Agrawal, Jason Hickey, and Nal Kalchbrenner. Metnet: A neural weather model for precipitation forecasting. *arXiv preprint arXiv:2003.12140*, 2020.
- Michail Tarasiou, Erik Chavez, and Stefanos Zafeiriou. Vits for sits: Vision transformers for satellite image time series. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10418–10428, 2023.
- Yunbo Wang, Mingsheng Long, Jianmin Wang, Zhifeng Gao, and Philip S Yu. Predrnn: Recurrent neural networks for predictive learning using spatiotemporal lstms. *Advances in neural information processing systems*, 30, 2017.
- Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, 2018.
- Wang-chun Woo and Wai-kin Wong. Operational application of optical flow techniques to radar-based rainfall nowcasting. *Atmosphere*, 8(3):48, 2017.

A APPENDIX

A.1 DETAILED BACKGROUND AND RELATED WORK

A.1.1 TASK CHALLENGE

Post-NWP optimization poses several distinct challenges that distinguish it from typical weather forecasting optimization problems and computer vision tasks.

Variable Selection and Modeling. In NWP, each pixel on the grid has various variables expressing the atmospheric feature state, which exhibit different statistical properties. This discrepancy includes spatial dependence and interdependence among variables, which violate the crucial assumption of identical and independently distributed data (Reichstein et al., 2019). The variables exhibit high correlation among themselves and also possess a degree of noise. Previous approaches have either used all available variables (Rojas-Campos et al., 2022) as input or relied on expert-based variable selection (Kim et al., 2022), which did not fully leverage the modeling capabilities.

Class Imbalance. The distribution of precipitation exhibits a significant imbalance, making model optimization challenging. A prior study (Shi et al., 2017) introduced WMSE, which assigned higher weighting factors to minority classes. Another study (Cao et al., 2022b) combined a reweighting loss with the MSE loss to mitigate the degradation in performance for majority classes. While these approaches have succeeded in improving forecast indicators for the minority class (heavy rainfall), they have inadvertently compromised the model’s performance on the majority class.

Lack of A Unified Benchmark. A previous study, KoMet (Kim et al., 2022), introduced a small dataset covering the time span of two years. Due to the limited data samples, models trained solely on such datasets may risk overfitting to specific data characteristics. Furthermore, KoMet only selected a subset of NWP variables as input. In contrast, another study (Rojas-Campos et al., 2022) utilized all 143 available NWP variables as input.

The limited size of the dataset, along with the lack of a standardized method for selecting variables, hinders research progress in improving the NWP post-processing task.

A.1.2 TASK FORMULATION

In this study, we consider optimizing the following model:

$$\min_{\mathbf{w}} \left\{ \mathcal{L}(\mathbf{w}; \mathcal{D}) \triangleq \mathbb{E}_{(X_t, y_t) \sim \mathcal{D}} [\ell(y_t; F(X_t, \mathbf{w}))] \right\} \quad (3)$$

where \mathcal{L} represents the objective function parameterized by \mathbf{w} on the dataset \mathcal{D} . As shown in Figure 6, the input is NWP predictions X_t , the corresponding ground-truth is rain observation y_t at time t , and ℓ denotes the loss function between the output of our proposed model $F(\cdot, \mathbf{w})$ and the ground-truth. The NWP predictions X_t are derived from the NWP model at time $t - L - \tau$, constituting a sequence denoted as $X_t = \mathbf{x}_{(t-L)}, \mathbf{x}_{(t-L+1)}, \dots, \mathbf{x}_{(t-2)}, \mathbf{x}_{(t-1)}$, where L signifies the sequence length and τ denotes the lead time. Our post-process model $F(\cdot, \mathbf{w})$ takes the sequence of NWP predictions X_t as input, aiming to predict a refined output \tilde{y}_t (at time t), where the rainfall observations y_t (at time t) serve as ground truth to train our model. In our multi-task framework, the prediction of our model at time t is defined as a classification forecast $\tilde{\mathbf{y}}_{cls}$ and a regression forecast $\tilde{\mathbf{y}}_{reg}$. Our proposed model $F(\cdot, \mathbf{w})$ is formulated as:

$$\tilde{\mathbf{y}}_{cls}, \tilde{\mathbf{y}}_{reg} = F(X_t, \mathbf{w}) \quad (4)$$

$$= F(\{\mathbf{x}_{(t-L)}, \mathbf{x}_{(t-L+1)}, \dots, \mathbf{x}_{(t-2)}, \mathbf{x}_{(t-1)}\}, \mathbf{w}) \quad (5)$$

where \mathbf{w} is the trainable parameters. Our model utilizes a classification head and a regression head to generate two final forecasts, $\tilde{\mathbf{y}}_{cls}$ and $\tilde{\mathbf{y}}_{reg}$. $\tilde{\mathbf{y}}_{cls}$ is a probability matrix and each item indicates the probability of a specific class among {‘non-rain’, ‘rain’, ‘heavy rain’}. $\tilde{\mathbf{y}}_{reg}$ is a prediction value of each pixel in the grid.

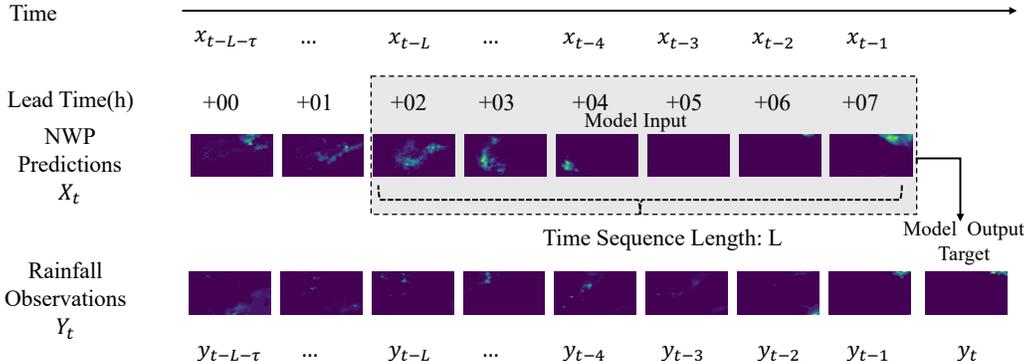


Figure 2: An illustrate of NWP post-processing task. NWP predictions X_t with a time sequence length of L is used as input, while rain observation y_t is used as ground truth.

Table 3: Comparison of PostRainBench’s three NWP datasets in different areas.

Area	Korea	Germany	China
Variable type		Pressure Level and Surface	
Variable numbers	12	143	28
Time period	2020-2021	2011-2018	2020-2021
Spatial resolution	12km \times 12km	2.8km \times 2.8km	1km \times 1km
Temporal resolution	1h	3h	3h
Temporal Window Size	6	1	1
Data shape (T C H W)	(6, 12, 50, 65)	(1, 143, 64, 64)	(1, 28, 64, 64)
Data split [train val test]	[4920, 2624, 2542]	[15189, 2725, 2671]	[2264, 752, 760]
Data size	47.9GB	16.2GB	3.6GB

A.2 DATASET DETAILS

A.2.1 POSTRAINBENCH DATASET SUMMARY

To address the issues of limited dataset size and the lack of a standardized criterion for variable selection, we introduce a unified benchmark comprising three datasets. Two of these datasets are sourced from prior research, while the third is collected from a public challenge. We describe our processing and standardization of the datasets below.

The first dataset, called KoMet (Kim et al., 2022), was collected in South Korea. The input data originates from GDAPS-KIM, a global numerical weather prediction model that furnishes hourly forecasts for diverse atmospheric variables. GDAPS-KIM operates at a spatial resolution of 12 km \times 12 km, resulting in a spatial dimension of 65 \times 50. The variables fall into two categories: pressure level variables and surface variables. For benchmarking purposes, 12 variables out of the 122 are selected according to Korean experts, and we follow this setting in our paper.

The second dataset originates from Germany (Rojas-Campos et al., 2022). This dataset covers the period from 2011 to 2018 and is confined to a selected area in West Germany. The input data is derived from the COSMO-DE-EPS forecast (Peralta et al., 2012), which provides 143 variables of the atmospheric state. For this dataset, the forecast with a 3-hour lead time is selected. A detailed description of the COSMO-DE-EPS output can be found in Schättler et al. (2008). The input data has a spatial resolution of 36 \times 36, while the output data is available at a resolution of 72 \times 72. To give a fair comparison between various algorithms, we perform interpolation on both to bring them to a consistent resolution of 64 \times 64.

The third dataset originates from China and provides hourly, 1 km \times 1 km resolution, 3-hour grid point precipitation data for the rainy season. This dataset spans from April to October in both 2020 and 2021. Additionally, it includes 3-hour lead time forecasts from a regional NWP model, with 28

surface and pressure level variables such as 2-meter temperature, 2-meter dew point temperature, 10-meter u and v wind components, and CAPE (Convective Available Potential Energy) values. For all variables provided, please refer to Table 4. Each time frame in this dataset covers a substantial spatial area, featuring a grid size of 430×815 . To maintain consistency, we interpolate this dataset to a more manageable 64×64 grid.

We summarize important details of the three datasets in the Table 3.

For Korea dataset and Germany dataset variables, please refer to previous research. For China dataset variables, please refer to Table 4.

A.2.2 CHINA DATASET VARIABLE

Table 4: List of variables contained in the China dataset.

Type	Long name	Short name	Level	Unit
Pressure Level	U-component of wind	u	200,500,700,850,925	(ms^{-1})
	V-component of wind	v	200,500,700,850,925	(ms^{-1})
	Temperature	T	500,700,850,925	(K)
	Relative humidity	rh liq	500,700,850,925	(%)
Surface	Rain	rain	*	(mm/h)
	Convective Rain	rain_thud	*	(mm/h)
	Large-scale Rain	rain_big	*	(mm/h)
	Convective Available Potential Energy"	cape	*	(J/kg)
	Precipitable Water	PWAT	*	(kg/m^2)
	Mean Sea Level	mssl	*	(hPa)
	2m temperature	t2m	*	$(^{\circ}C)$
	2m dew point temperature	d2m	*	$(^{\circ}C)$
	10m component of wind	u10m	*	(ms^{-1})
	10m v component of wind	v10m	*	(ms^{-1})

A.2.3 DATA DISTRIBUTION

We analyze the distribution of the observed precipitation data, which serves as the ground truth, across the three datasets. In accordance with the framework outlined in Kim et al. (2022), we categorize precipitation into two types: rain and heavy rain, each with its set of evaluation metrics and frame this forecasting problem as a three-class classification task. It is important to note that the threshold for defining heavy rain can vary by location due to differences in rainfall frequency influenced by geographical and climatic factors.

In Germany dataset, Rojas-Campos et al. (2022) explores various thresholds including 0.2, 0.5, 1, 2, and 5, we adopt a rain threshold of 10^{-5} mm/h since its distribution is concentrated in $[0,1]$ and we adhere to the rain threshold of 0.1mm/h adopted by Kim et al. (2022) In Korea dataset, we adhere previous heavy rain threshold of 10mm/h and opt for a unified threshold of 2mm/h in another two datasets, enabling a more equitable comparison. The distribution and the rain categorization of the three datasets are presented in Table 5.

It is evident that all three datasets exhibit significant imbalances, which presents a great challenge to predict extreme weather scenarios.

A.3 MODEL DETAILS

A.3.1 BASELINES

U-Net(Ronneberger et al., 2015) is a model specifically crafted to address the challenge of image segmentation in biomedical images. It excels in capturing essential features in a reduced-dimensional form during the propagation phase of its encoder component.

Table 5: Statistics of three datasets.

Dataset	Rain rate (mm/h)	Proportion (%)	Rainfall Level
KoMet	[0.0, 0.1)	87.24	No Rain
	[0.1, 10.0)	11.57	Rain
	[10.0, ∞)	1.19	Heavy Rain
Germany	[0.0, 10^{-5})	85.10	No Rain
	[10^{-5} , 2.0)	13.80	Rain
	[2.0, ∞)	1.10	Heavy Rain
China	[0.0, 0.1)	91.75	No Rain
	[0.1, 2.0)	3.81	Rain
	[2.0, ∞)	4.44	Heavy Rain

ConvLSTM (Shi et al., 2015; 2017) is a hybrid model integrating LSTM and convolutional operations. LSTMs are tailored for capturing temporal relationships, while convolutional operations specialize in modeling spatial patterns. This combination allows ConvLSTM to effectively model both temporal and spatial relationships within sequences of images.

MetNet (Sønderby et al., 2020) incorporates a spatial downsampler, achieved through convolutional layers, to reduce input size. Its temporal encoder employs the ConvLSTM structure, enabling the capture of spatial-temporal data on a per-pixel basis. The feature map subsequently undergoes self-attention in the Spatial Aggregator to integrate global context, before being processed by a classifier that outputs precipitation probabilities for each pixel.

Swin-Unet (Cao et al., 2022a) is a Unet-like Transformer. The tokenized image patches are fed into the Swin Transformer-based U-shaped Encoder-Decoder architecture with skip connections for local-global semantic feature learning. Specifically, it uses hierarchical Swin Transformer (Liu et al., 2021) with shifted windows as the encoder and decoder. The overall architecture of Swin-Unet is presented in Figure 3. In our multi-task framework, two linear projection layers are applied to output the pixel-level classification and regression predictions. **ViT** (Dosovitskiy et al., 2020) apply a pure Transformer architecture on image data, by proposing a simple, yet efficient image tokenization strategy. We follow previous work (Tarasiou et al., 2023) to employ Transformers for dense prediction.

FourCastNet (Pathak et al., 2022) is a data-driven global weather forecasting model known for its rapid and accurate predictions, excelling in high-resolution forecasting of complex meteorological variables, which is based on Adaptive Fourier Neural Operators (AFNO).

A.3.2 SWIN-UNET ARCHITECTURE

The overall architecture of Swin-Unet is presented in Figure 3. In our multi-task framework, two linear projection layers are applied to output the pixel-level classification and regression predictions.

A.3.3 CHANNEL ATTENTION MODULE

CAM aggregates spatial information of a feature map by using both average-pooling and max-pooling operations, generating two different spatial context descriptors: $\mathbf{F}_{\text{avg}}^c$ and $\mathbf{F}_{\text{max}}^c$. Both descriptors are forwarded to a shared multi-layer perceptron (MLP) to produce a channel attention map $\mathbf{M}_c \in \mathbb{R}^{C \times 1 \times 1}$. To reduce parameter overhead, the hidden activation size is set to $\mathbb{R}^{C/r \times 1 \times 1}$, where r is the reduction ratio. After the shared network, the two output feature vectors are merged with element-wise summation. We employ a residual connection (He et al., 2016) by adding the attention map to the original input, which serves as the input for the subsequent backbone stage. In short, the channel attention is computed as:

$$\begin{aligned} \mathbf{M}_c(\mathbf{F}) &= \sigma(\text{MLP}(\text{AvgPool}(\mathbf{F})) + \text{MLP}(\text{MaxPool}(\mathbf{F}))) \\ &= \sigma(\mathbf{W}_1(\mathbf{W}_0(\mathbf{F}_{\text{avg}}^c)) + \mathbf{W}_1(\mathbf{W}_0(\mathbf{F}_{\text{max}}^c))), \end{aligned} \tag{6}$$

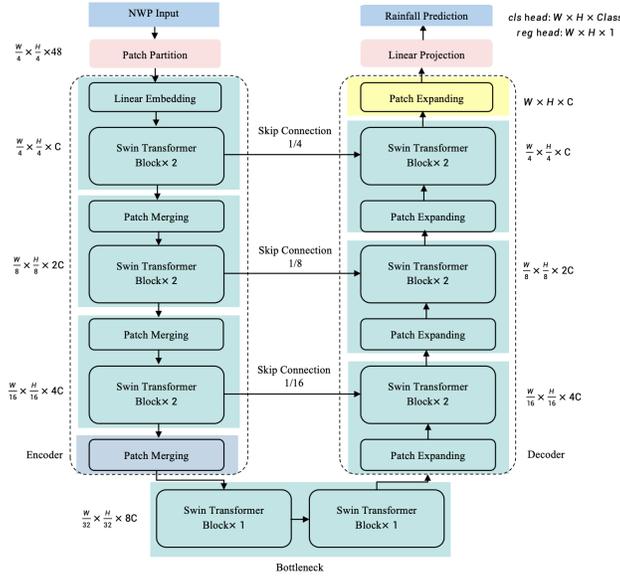


Figure 3: The architecture of Swin-Unet, which is composed of encoder, bottleneck, decoder and skip connections. Encoder, bottleneck and decoder are all constructed based on swin transformer block.

where σ denotes the sigmoid function, $\mathbf{W}_0 \in \mathbb{R}^{C/r \times C}$, and $\mathbf{W}_1 \in \mathbb{R}^{C \times C/r}$. We choose $r = 16$. Note that the MLP weights, \mathbf{W}_0 and \mathbf{W}_1 , are shared for both inputs and the activation function is followed by \mathbf{W}_0 . We choose GeLU activation function instead of ReLU.

The resulting feature maps are then input to the Swin-Unet backbone, as shown in Figure 1.

The backbone model is connected to a classification head and a regression head, which are learned under our proposed multitask learning framework as described in the next section.

A.4 EVALUATION METRICS

In terms of evaluation, we adopt commonly used multi-class classification metrics for precipitation forecasting by previous works (Kim et al., 2022). The evaluation metrics are calculated based on the number of true positives (TP_k), false positives (FP_k), true negatives (TN_k), and false negatives (FN_k) for some generic class k . We describe all the metrics we consider as follows:

- **Critical Success Index (CSI)** (Donaldson et al., 1975) is a categorical metric that takes into account various elements of the confusion matrix, similar with F1-score having the value as $\frac{TP_k}{TP_k + FN_k + FP_k}$.
- **Heidke Skill Score (HSS)** (Woo & Wong, 2017) as stated by (Hogan et al., 2010), is more equitable in evaluating the forecasting performance. Higher HSS means better performance and a positive HSS indicates that a forecast is better than a random-based forecast. HSS is calculated as $\frac{2 \times (TP_k \times TN_k - FN_k \times FP_k)}{FP_k^2 + TN_k^2 + 2 \times TP_k \times FN_k + (FP_k + TN_k)(TP_k + FP_k)}$.
- **Accuracy (ACC)** provides a comprehensive assessment of how accurately the model predicts outcomes across the entire dataset.
- **Probability of Detection (POD)** is a recall calculated as $\frac{TP_k}{TP_k + FP_k}$.
- **False Alarm Ratio (FAR)** (Barnes et al., 2009) represents the number of false alarms in relation to the total number of warnings or alarms, indicating the probability of false detection. It is computed as $\frac{FN_k}{TP_k + FN_k}$.
- **Bias** quantifies the ratio between the observed frequency of a phenomenon and the frequency predicted by the forecasting model. $\frac{TP_k + FP_k}{TP_k + FN_k}$. If the value is greater than 1, it

signifies that the forecast model predicts the occurrence more frequently than the actual phenomenon. Consequently, a bias value closer to 1 indicates a more accurate forecast.

A.5 TRAINING DETAILS

The datasets are split into training, validation, and test sets following the configurations outlined in previous studies. For the China dataset, we randomly partition the data into a 6:2:2 ratio. To ensure consistency with prior studies, we select the model with the best CSI performance on the validation set and report its performance on the test set. Each model is run with three different random seeds for robust performance. We use the Adam optimizer for all models.

For the Korea dataset, baseline models are trained with a learning rate of 0.001 (as mentioned in Kim et al. (2022)), while Swin-Unet models are trained with a learning rate of 0.0001. Consistent with previous settings, a batch size of 1 is employed, and all models are trained for 50 epochs. We apply a weight of [1, 5, 30] for the CE Loss. We utilize a hyperparameter α of 100 for the MSE Loss on all datasets. For the Germany dataset, baseline models are trained with a learning rate of 0.001 (as mentioned in Rojas-Campos et al. (2022)), whereas Swin-Unet models are trained with a learning rate of 0.0001. The batch size remains consistent with previous settings at 20, and all models are trained for 30 epochs. We utilize a class weight of [1, 5, 30]. For the China dataset, all models are trained with a learning rate of 10^{-4} for 100 epochs. The weight configuration used is [1, 15, 10].

A.6 ADDITIONAL EXPERIMENTS

A.6.1 EXPERIMENTS WITH MORE METRICS

We report more evaluation metrics of all models in this section.

Table 6: Evaluation metrics on three datasets. Best performances are marked in **bold**. '↑' indicates that higher scores are better, '↓' indicates that higher scores are worse.

		Rain						Heavy Rain					
		Acc↑	POD↑	CSI↑	FAR↓	Bias	HSS↑	Acc↑	POD↑	CSI↑	FAR↓	Bias	HSS↑
Korea	NWP	0.747	0.633	0.263	0.690	2.042	*	0.985	0.055	0.045	0.795	0.266	*
	U-Net	0.860	0.430	0.305	0.489	0.841	0.387	0.987	0.001	0.001	0.750	0.002	0.001
	ConvLSTM	0.860	0.446	0.312	0.492	0.878	0.395	0.986	0.011	0.010	0.874	0.083	0.018
	MetNet	0.853	0.457	0.307	0.517	0.946	0.384	0.987	0.013	0.012	0.805	0.067	0.023
	Ours	0.832	0.559	0.322	0.569	1.299	0.388	0.979	0.067	0.048	0.908	0.729	0.068
Germany	NWP	0.728	0.925	0.338	0.652	2.657	0.252	0.980	0.434	0.178	0.767	1.863	0.173
	U-Net	0.903	0.631	0.495	0.305	0.908	0.605	0.990	0.053	0.051	0.412	0.090	0.095
	ConvLSTM	0.896	0.623	0.475	0.334	0.935	0.583	0.990	0.048	0.045	0.566	0.111	0.085
	MetNet	0.895	0.653	0.483	0.349	1.003	0.590	0.990	0.000	0.000	0.694	0.001	0.001
	Ours	0.884	0.811	0.513	0.418	1.393	0.610	0.989	0.280	0.207	0.557	0.632	0.338
China	NWP	0.843	0.433	0.164	0.792	2.082	0.123	0.903	0.348	0.110	0.861	2.512	0.089
	U-Net	0.914	0.071	0.060	0.725	0.261	0.084	0.950	0.053	0.042	0.821	0.294	0.064
	ConvLSTM	0.909	0.083	0.066	0.756	0.339	0.088	0.941	0.099	0.066	0.837	0.607	0.094
	MetNet	0.915	0.086	0.072	0.680	0.268	0.106	0.947	0.104	0.076	0.778	0.466	0.118
	Ours	0.873	0.454	0.216	0.708	1.553	0.289	0.943	0.210	0.135	0.727	0.768	0.209

For accuracy (Acc), our model performs lower than the baseline deep learning models but higher than NWP. However, it's important to note that accuracy may not provide realistic insights in an extremely imbalanced case. If the model predicts all instances as no-rain, it could achieve a better score. For probability of detection (Pod), our model ranks second only to NWP and outperforms all deep learning models. In terms of critical success index (CSI) and Heidke skill score (HSS), our model consistently outperforms the baseline models, as discussed earlier. The false alarm ratio (Far) measures whether the forecasting model predicts an event more frequently than it actually occurs. Our model exhibits higher but acceptable values in the rain category compared to other deep-learning

models, reflecting the trade-off between enhanced forecasting ability and overforecast. In the heavy rain category, our model’s bias is less than 1 and closer to 1, indicating a more accurate forecast.

A.6.2 COMPARISON WITH FOURCASTNET

For the Korea dataset, our model exhibits superior performance to FourCastNet in both rain CSI and heavy rain CSI metrics, with a marginal shortfall in rain HSS, where it trails by 1.8% behind FourCastNet. It is important to highlight that FourCastNet’s predictive capability does not surpass that of NWP algorithms for heavy rain scenarios.

Regarding the Germany dataset, our model demonstrates an advancement over FourCastNet in all metrics for both rain and heavy rain, whereas FourCastNet does not demonstrate an advantage over NWP algorithms in heavy rain predictions.

For the China dataset, our model demonstrates comprehensive outperformance across all metrics when compared to FourCastNet. While FourCastNet posts a modest 3.6% gain over NWP methods in heavy rain forecasting, our approach achieves a substantial 31.8% improvement, marking a significant enhancement in predictive accuracy.

Table 7: Experiment result compared with FourCastNet. Each model undergoes three runs with different random seeds, and we report the mean, standard deviation (std), and best performance in terms of CSI and HSS. The best results are highlighted in **bold**. In the context of the results, ‘↑’ indicates that higher scores are better.

		Rain				Heavy Rain			
		CSI↑		HSS↑		CSI↑		HSS↑	
		Mean(Std)	Best	Mean(Std)	Best	Mean(Std)	Best	Mean(Std)	Best
Korea	NWP	0.263(±0.000)		*		0.045(±0.000)		*	
	FourCastNet	0.314 (±0.016)	0.325	0.391 (±0.023)	0.409	0.011(±0.008)	0.017	0.020(±0.014)	0.029
	Ours	0.321 (±0.005)	0.326	0.384(±0.007)	0.389	0.052 (±0.010)	0.058	0.089 (±0.017)	0.097
Germany	NWP	0.338(±0.000)		0.252(±0.000)		0.178(±0.000)		0.173(±0.000)	
	FourCastNet	0.494 (±0.009)	0.504	0.595(±0.009)	0.601	0.157(±0.034)	0.185	0.265(±0.051)	0.306
	Ours	0.514 (±0.003)	0.518	0.609 (±0.006)	0.616	0.209 (±0.014)	0.224	0.339 (±0.020)	0.359
China	NWP	0.164(±0.000)		0.123(±0.000)		0.110 (±0.000)		0.089(±0.000)	
	FourCastNet	0.163 (±0.006)	0.167	0.219(±0.010)	0.230	0.114(±0.013)	0.129	0.166(±0.023)	0.192
	Ours	0.208 (±0.007)	0.216	0.274 (±0.014)	0.289	0.145 (±0.015)	0.163	0.225 (±0.019)	0.246

* For Korea dataset, NWP method’s HSS is not reported. For all NWP method, we only have the mean value.

A.6.3 ABLATION STUDY ON BACKBONE

We conduct another ablation study by replacing Swin-Unet backbone with ViT (Dosovitskiy et al., 2020) backbone under our CAMT framework in Table 8.

For the Korea dataset, ViT outperforms Swin-Unet in rain CSI and HSS but shows a slight decrease in heavy rain CSI. Importantly, its performance remains higher than that of NWP, which shows the effectiveness of CAMT. For the Germany dataset, though its performance on rain CSI is limited, the ViT model still demonstrates a remarkable performance in heavy rain CSI and surpasses NWP. For the China dataset, ViT outperforms all baseline models and is only second to Swin-Unet.

These experiments highlight the potential of the ViT model. We also conduct experiments with three baseline models but observe limited improvements. We believe that addressing the challenge of imbalanced precipitation forecasting requires a more robust backbone and the use of our CAMT framework, which incorporates multi-task information to enrich the learning process of this task.

A.7 QUANTITATIVE ANALYSIS

A.7.1 PERFORMANCE ON DIFFERENT LEAD TIME ON KOREA DATASET

As shown in Figure 4, within the lead time interval of 6 to 20, we observe that the CSI for rain reaches a peak at a lead time of 10 before exhibiting a declining trend, whereas the CSI for heavy rain peaks at a lead time of 9, subsequently showing a fluctuating trajectory.

Table 8: Ablation study with ViT backbone, we highlight the best results in **bold**.

		Rain				Heavy Rain			
		CSI \uparrow		HSS \uparrow		CSI \uparrow		HSS \uparrow	
		Mean(Std)	Best	Mean(Std)	Best	Mean(Std)	Best	Mean(Std)	Best
Korea	ViT+CAMT	0.326 (± 0.004)	0.329	0.394 (± 0.001)	0.395	0.049 (± 0.010)	0.055	0.083 (± 0.017)	0.097
	Swin-Unet+CAMT	0.321 (± 0.005)	0.326	0.384 (± 0.007)	0.389	0.052 (± 0.010)	0.058	0.089 (± 0.017)	0.097
Germany	ViT+CAMT	0.484 (± 0.004)	0.488	0.576 (± 0.005)	0.581	0.194 (± 0.023)	0.041	0.050 (± 0.043)	0.078
	Swin-Unet+CAMT	0.514 (± 0.003)	0.518	0.609 (± 0.006)	0.616	0.209 (± 0.014)	0.224	0.339 (± 0.020)	0.359
China	ViT+CAMT	0.177 (± 0.004)	0.181	0.217 (± 0.006)	0.224	0.068 (± 0.033)	0.105	0.091 (± 0.052)	0.149
	Swin-Unet+CAMT	0.208 (± 0.007)	0.216	0.274 (± 0.014)	0.289	0.145 (± 0.015)	0.163	0.225 (± 0.019)	0.246

Expanding the analysis to a lead time range of 6 to 87, both rain and heavy rain CSI exhibit parallel trends, with heavy rain demonstrating superior performance over extended lead times, likely reflective of inherent data characteristics. Across all evaluated lead times from 6 to 87, our model’s mean performance is enhanced, underscoring the comprehensive superiority of our modeling approach.

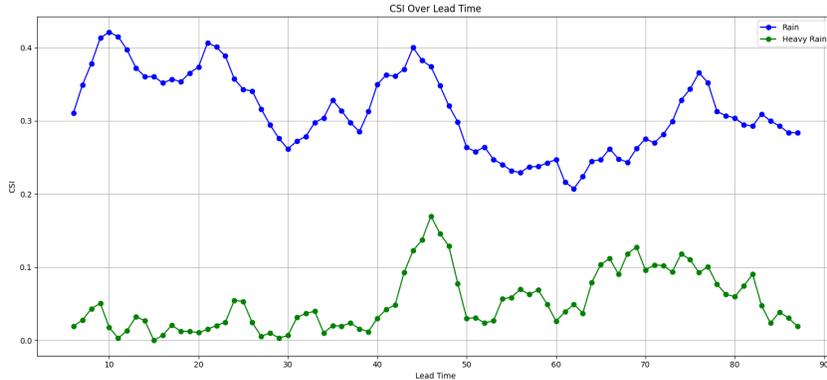


Figure 4: CSI scores of Korea Dataset for rain and heavy rain classification with lead times ranging from 6 to 87 hours.

A.7.2 VALIDATION LOSS ON GERMANY DATASET

In our ablation study, we visualized the validation loss for different configurations of our model on the Germany Dataset to assess the impact of each proposed component. The validation loss curve for the standalone Swin-Unet displayed an upward trend, suggesting a potential for overfitting or an insufficient capture of the dataset’s essential patterns. Conversely, the integration of our proposed Channel Attention Module (CAM) and Weighted Loss (WL) resulted in a downward trend of the loss over epochs, indicating effective learning of the data distribution and improved generalizability of the model.

The CAM, with its targeted focus on salient features, and the WL, which addresses class imbalance, have shown a discernible positive influence on the model’s learning process, as demonstrated by a consistent reduction in validation loss. This reduction substantiates our method’s capability to tackle the specific challenges associated with precipitation forecasting in imbalanced datasets.

Ultimately, the depicted loss curves validate our method’s proficiency in grasping the complexities of the forecasting task, where the integrated components not only counteract overfitting but also significantly bolster the model’s forecasting accuracy.

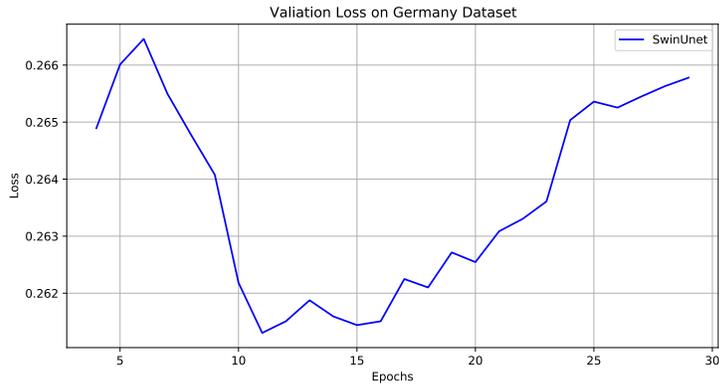


Figure 5: Valiation loss on Germany Dataset with Swin-Unet.

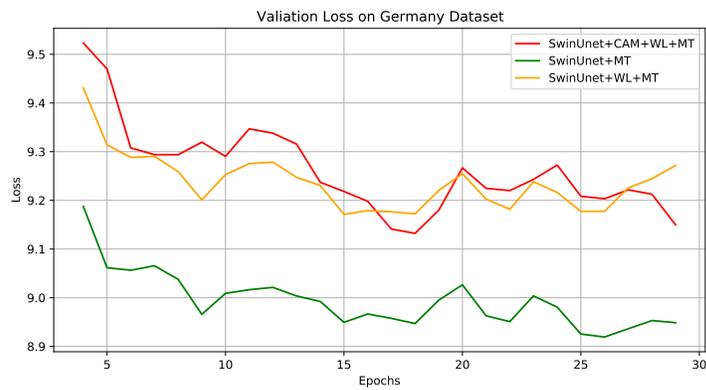


Figure 6: Valiation loss on Germany Dataset with Swin-Unet and proposed components: CAM (Channel Attention Module), WL (Weighted Loss), and MT (Multi-task Learning).