Explainable AI-Based Interface System for Weather Forecasting Model *

Soyeon Kim[1][0009−0001−5037−0902], Junho Choi[1][0000−0002−7800−6950], Yeji Choi[2][0000−0002−8212−1126], Subeen Lee[1][0009−0001−7996−4114], Artyom Stitsyuk[1][0009−0005−6446−137X], Minkyoung Park[3][0009−0004−9420−4406], Seongyeop Jeong[1][0009−0008−8480−8117], Youhyun Baek[4][0000−0001−6362−4353], and Jaesik Choi[1,5][0000−0002−4663−3263]

[1] Korea Advanced Institute of Science and Technology(KAIST), Daejeon, Korea
soyeon.k, junho.choi, forestsoop, stitsyuk, jrneomy, seongyeop.jeong@kaist.ac.kr
[2] SI-Analytics, Daejeon, Korea
yejichoi@si-analytics.ai
[3] National Institute of Meteorological Sciences(NIMS), Jeju, 63568, Korea
yhbaek88@korea.kr
[4] INEEJI, Gyeonggi, Korea
jaesik.choi@kaist.ac.kr

Abstract. Machine learning (ML) is becoming increasingly popular in meteorological decision-making. Although the literature on explainable artificial intelligence (XAI) is growing steadily, user-centered XAI studies have not extend to this domain yet. This study defines three requirements for explanations of black-box models in meteorology through user studies: statistical model performance for different rainfall scenarios to identify model bias, model reasoning, and the confidence of model outputs. Appropriate XAI methods are mapped to each requirement, and the generated explanations are tested quantitatively and qualitatively. An XAI interface system is designed based on user feedback. The results indicate that the explanations increase decision utility and user trust. Users prefer intuitive explanations over those based on XAI algorithms even for potentially easy-to-recognize examples. These findings can provide evidence for future research on user-centered XAI algorithms, as well as a basis to improve the usability of AI systems in practice.

Keywords: User-Centered Explainable AI · Interactive Visualization · Feature Attribution · Precipitation Forecasting

* Supported by the Korean Institute of Information & Communications Technology Planning & Evaluation (IITP) and the Korean Ministry of Science and ICT(MSIT) under grant agreement No. 2019-0-00075 (Artificial Intelligence Graduate School Program (KAIST)) and No. 2022-0-00984 (Development of Plug-and-Play Explainable Artificial Intelligence Method), and from the Korea Meteorological Administration (KMA) and Korean National Institute of Meteorological Sciences (NIMS) under grant agreement No. KMA2021-00123 (Developing Intelligent Assistant Technology and Its Application for Weather Forecasting Process).

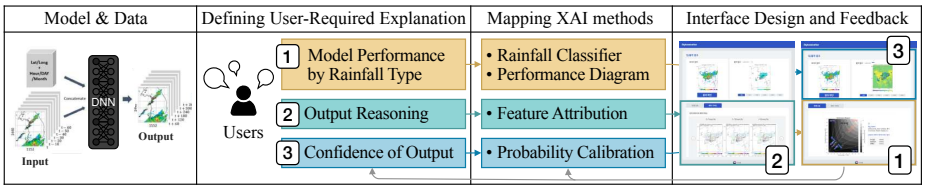

Fig. 1. Workflow for developing a user-centered explainable artificial intelligence(XAI) interface system. The system is developed based on the procedures established in the previous literature [19, 20]. The scope of explanations is defined based on the requirements set by the practitioners; appropriate XAI algorithms are selected based on the defined scope; and the interface is designed with user feedback

1 Introduction

Weather prediction has always been an integral part of human society due to its significant socioeconomic impact, influencing "agricultural output, industrial output, labor productivity, energy demand, health, conflict, and economic growth among other outcomes" [8] as well as ecosystems and ecosystem services [11]. With the increasing volatility of meteorological patterns caused by the climate crisis, economic losses from extreme weather events are on a rapid incline [24, 41]. Accurate weather forecasting is crucial for mitigating the effects of these scenarios.

Operational weather forecasting is conventionally performed through Numerical Weather Prediction (NWP), a process of simulating future weather patterns using a comprehensive set of equations that describe the physical dynamics of the atmosphere [15]. Although it has a long history and sees use even today, NWP faces several challenges such as high computational costs and sensitivity to the derived initial conditions [31]. Data-driven deep learning models for weather prediction are seen as a potential alternative, being able to exploit the growing availability of weather data and make predictions for a fraction of the cost of operating NWP models [29].

One issue faced by practitioners in producing weather forecasts is the vast amount of documents required to produce the forecasts. For example, Korea Meteorological Agency (KMA) creates 2.2TB worth of data daily on average for weather forecasts [18]. The sheer size of the data can be extremely burdensome for the forecasters, who not only have limited time when making short-term forecasts and associated decision-making, but also need to continuously monitor the occurrence of sudden extreme weather patterns. One of the reasons for requiring large data lies with the difficulty in accurate prediction of rainfall. If the accuracy of rainfall prediction can be improved through the use of deep learning, it could reduce some of the burden placed on the forecasters so that their efforts could be invested elsewhere.

A key issue preventing the use of deep learning models in operational forecasting is their lack of interpretability [31]. While the state-of-the-art models [10, 16, 30, 38] may make accurate predictions, they tend to be black boxes

– a user cannot determine how the models infer these outcomes. A forecaster would not be able to accept predictions without sufficient justifications due to the high stakes associated with wrong predictions. The extensive array of techniques in the field of explainable artificial intelligence (XAI) can help meet these requirements [1, 12, 34]; unfortunately, the sheer number of available techniques makes it difficult to determine which methods should be used. One potential approach of filtering the appropriate techniques is to center the explanations around its intended audience. An appropriate explanation is dependent on the task performed by a model and the audience of the explanation [21, 25]. Therefore, an explanation system should be centered around its users, regardless of domain. A recent study of the user-explained AI (UXAI) [6] even claims that users may not be satisfied by an explanation that has considered the users in its design if it has not been made *with* the users. Despite the increasing interest in both user-centric [21] and regular XAI in the meteorological domain [3, 22, 23], there seems to be a distinct lack of user-centered XAI studies in meteorology. This paper attempts to fill this gap by following a user-centered XAI framework to create a prototype system that explains a precipitation prediction AI model. Specifically, the paper follows the process described by [19] and [20]: (a) the scope of explanations is defined through an XAI question bank, which divides the typical questions that could be asked by a user into several major categories, (b) appropriate XAI methods are selected based on the categories that the questions belong to, and (c) an interface system is designed based on user input and feedback to express the explanations.

The main contributions of this paper are as follows:

- Demonstrates the procedures of the user-centric XAI development framework from an operational perspective.
- Creates a user-experience-based prototype of the XAI system in the meteorological domain.
- Analyzes the available XAI methods and discusses their practical limitations.

The eventual objective of our work is to provide accurate and trustworthy information required by the user as an end-product of a single map, reducing the procedural burden shouldered by the forecasters in the current system.

2 Materials

2.1 Model and Data

The aim of this study is to design a user-centered interface system for explaining UNet2, a UNet-based model (an unpublished variant of DeepRaNE [17]) developed by the National Institute of Meteorological Sciences (NIMS) for 2020 radar synthesis data for very short-term rainfall intensity prediction (Figure 2). UNet2 consists of a denoising autoencoder followed by a convolutional neural network-based U-Net architecture and addresses a segmentation task of predicting three rainfall intensity intervals (no rain 0-1 mm/hr, light rain 1-10 mm/hr, and heavy

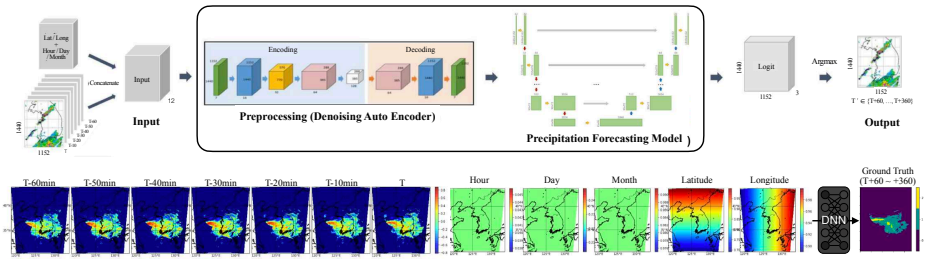

Fig. 2. The target precipitation forecasting model and data. The data consists of radar hybrid scan reflectivity.

rain 10 mm/hr over) between 1 and 6 hours in the future at 1-hour intervals. The class intervals have been established by domain experts. The input data consists of seven radar data sequence at ten minutes intervals, two spatial features for longitude and latitude, and three temporal features for year, month, and day of the current date. The data are concatenated into 12 channels following an early fusion scheme. The performance of UNet2 is comparable to the MetNet [38] and HRRR numerical models for very short-term predictions(Figure 3). Specifically, when predicting rainfall for a lead time of 1 hour and rainfall rates between 1-10mm/hr, UNet2 and MetNet have F1 scores of 0.824 and 0.822, respectively. For heavy rainfall rates over 10mm/hr, UNet2 and MetNet have F1 scores of 0.604 and 0.480, respectively.

Model	Score	Class (mm/hr)	Lead Time (hour)					
			1	2	3	4	5	6
Unet1	F1	1	0.772	0.692	0.642	0.595	**0.570**	**0.519**
		10	0.558	0.398	0.315	**0.265**	0.206	0.121
	CSI	1	0.629	0.629	0.472	0.424	**0.398**	**0.351**
		10	0.387	0.249	0.187	**0.153**	**0.115**	0.065
Unet2 (Target)	F1	1	**0.824**	0.710	0.642	0.579	0.514	0.462
		10	**0.604**	**0.419**	0.326	0.243	0.174	0.126
	CSI	1	**0.701**	0.551	0.473	0.407	0.346	0.300
		10	**0.433**	0.265	0.195	0.139	0.095	0.067
MetNet	F1	1	0.822	**0.721**	**0.665**	**0.597**	0.547	0.502
		10	0.480	**0.419**	**0.328**	0.255	**0.207**	**0.168**
	CSI	1	0.697	**0.697**	**0.487**	**0.425**	0.376	0.335
		10	0.316	**0.316**	**0.196**	0.146	**0.115**	**0.092**

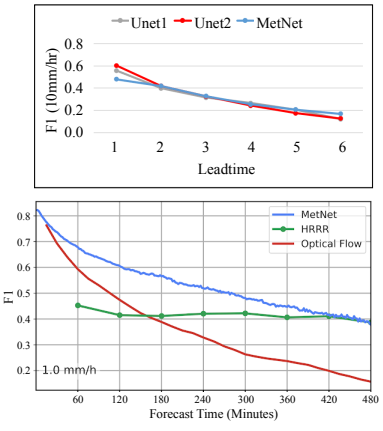

Fig. 3. The performance of the target model. UNet1 and UNet2 built by NIMS are comparable to MetNet [38] and HRRR numerical model for very short-term predictions.

3 Methods

3.1 User Requirements of Explanation

This study has been performed with discussions from sixteen online meetings with NIMS, as well as three in-person external advisories from domain experts from 27 April 2022 to 12 April 2023.

User Study. An XAI question bank [19, 20] is utilized in the early phase of interviews to brainstorm the desired explanations from AI systems. Based on the discussion, the user requirements can be stated as follows. First, forecasters are interested in the consistency of the model inferences in various rainfall situations. If systematic biases for each rainfall type are provided, it can help the forecasters decide whether to use the model in practice. Second, forecasters consider the movement, growth, and dissipation of the convection cell as key factors for predicting the change of very short-term precipitation clouds around a 6-hour scale. In particular, they would like to identify the precursors to pinpoint the seeds that are the most susceptible to convective system development. Through the precursors, the users can indicate the locations that require more intensive monitoring. Finally, the users are interested in the local reliability of the predictions. For the rest of this study, these three requirements are referred to as model performance explanation by rainfall type, output reasoning explanation, and confidence of output explanation, respectively.

Mapping XAI methods. Appropriate XAI methods are selected to address each need. First, a rainfall type classifier is combined with performance diagram for each rainfall type for generating a model performance explanation (Section 3.2). Second, feature attribution is used for output reasoning explanation since the associated techniques can evaluate the contributions of the input features for generating the predictions (Section 3.3). Lastly, a probability calibration technique is adopted for model confidence explanation (Section 3.4).

3.2 Explanation 1: Model Performance by Rainfall Types

Rainfall Type Classifier. For this explanation, an input sample is assigned to a rainfall category using a deep learning classifier; then, the model's predictive performance for the corresponding rainfall type is analyzed. This setup allows for a comparison of model performance over different rainfall scenarios.

The rainfall type classifier is built by fine-tuning the parameters from the pre-trained encoder of the target model. Self-organizing map (SOM)-based rainfall type classification data and its quantitative labels provided by NIMS based on the characteristics of the Korean Peninsula have been used for the experiment. The five rainfall types are monsoon front (southern region), monsoon front (central region), isolated thunderstorm, extratropical cyclone (east coast), and extratropical cyclone (inland). These precipitation types are often used by

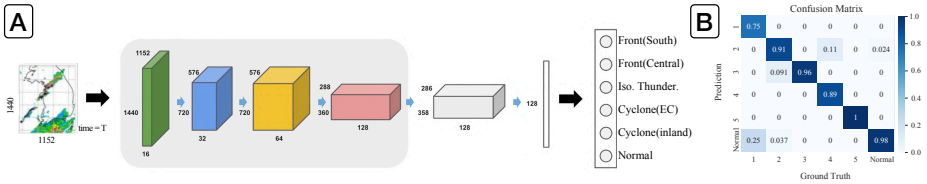

Fig. 4. The structure of the precipitation classifier(A) and the resulting confusion matrix(B). The rainfall types are based on a SOM-based weather classification study (an unpublished result of [36] with the same research procedure on a specific region).

forecasters in practice. 29, 280, 53, 43, and 24 samples are used for each of the five types of rainfall in 2020. Additionally, 218 cases are sampled in equal intervals for the no-rain type. The data is divided into 60% for training, 20% for validation, and 20% for testing. For the training dataset, a sampler that follows a multinomial distribution using the probability parameter as the inverse of the number of samples of each class in the dataset is used to solve the class imbalance problem. The classifier is optimized using Adam solver with a learning rate of 1e-6 and weight decay of 1e-8. Weighted cross-entropy loss is adopted to adjust for the classes having deficient samples. The classifier performance shows an accuracy of 93.07%.

Performance Diagram. The performance diagram is a method of visualizing the overall performance of a model [32] and can express important model evaluation indicators in the meteorological domain such as bias, critical success index(CSI), probability of detection(POD), and success ratio in a single chart(Figure 6. To alleviate the problem of imbalanced rainfall intensities, where the rainfall amounts of interest infrequently occur in the real world, the metrics are computed for the light rainfall intensity and more(1 mm/hr over) and the heavy rainfall intensity(10 mm/hr over) as shown in Figure 5 and are averaged. Formally,

$$\text{ModifiedPOD} = \frac{1}{2}\left(\frac{\text{Hit}_{1\ \text{(mm/hr) over}}}{\text{Hit}_{1\ \text{over}} + \text{Miss}_{1\ \text{over}}} + \frac{\text{Hit}_{10\ \text{over}}}{\text{Hit}_{10\ \text{over}} + \text{Miss}_{10\ \text{over}}}\right) \quad (1)$$

$$\text{ModifiedFAR} = \frac{1}{2}\left(\frac{\text{FalseAlarm}_{1\ \text{(mm/hr) over}}}{\text{FalseAlarm}_{1\ \text{over}} + \text{Hit}_{1\ \text{over}}} + \frac{\text{FalseAlarm}_{10\ \text{over}}}{\text{FalseAlarm}_{10\ \text{over}} + \text{Hit}_{10\ \text{over}}}\right) \quad (2)$$

$$\text{ModifiedF1} = \frac{1}{2}\left(\frac{\text{Hit}_{1\ \text{(mm/hr) over}}}{\text{Hit}_{1\ \text{over}} + \frac{1}{2}(\text{Miss}_{1\ \text{over}} + \text{FalseAlarm}_{1\ \text{over}})} + \frac{\text{Hit}_{10\ \text{over}}}{\text{Hit}_{10\ \text{over}} + \frac{1}{2}(\text{Miss}_{10\ \text{over}} + \text{FalseAlarm}_{10\ \text{over}})}\right) \quad (3)$$

		True (mm/hr)		
		0-1	1-10	10~
Predict	0-1	TN1 (Correct reject)	FN1 (Miss)	
	1-10	FP1 (False alarm)	TP1 (Hit)	
	10~			

		True (mm/hr)		
		0-1	1-10	10~
Predict	0-1	TN2 (Correct reject)		FN2 (Miss)
	1-10			
	10~	FP2 (False alarm)		TP2 (Hit)

Fig. 5. Confusion matrices to calculate performance metrics on the imbalanced data.

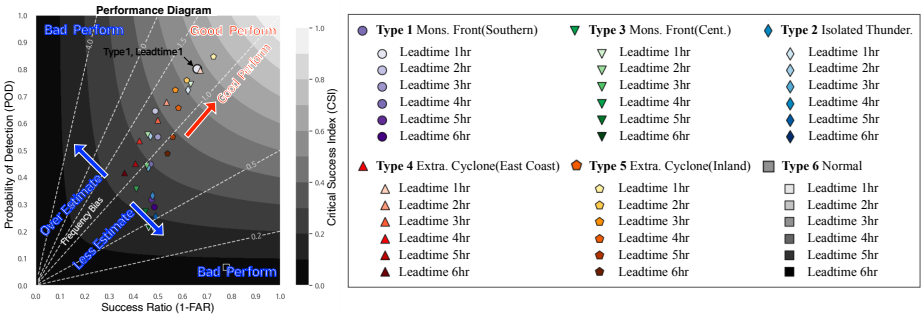

Fig. 6. Performance diagram. The diagram helps visualize the overall performance of bias, CSI, POD, and success ratio in a single chart.

As results in Figure 6, the performance diagram shows that for a lead time of 1 hour, the model has the best performance for rainfall type 5 - inland extratropical cyclone. The model is a little overestimated overall, but less estimated on the long lead times. The worst performance arises in the type of normal weather at the lead time of 6 hours. The low POD suggests that the model fails to predict the real rainfall at this lead time.

3.3 Explanation 2: Output Reasoning

Feature attribution methods analyze the contribution of the inputs for a model's prediction. As shown in Figure 7 feature attribution methods allow users to investigate the reason why the model infers the development or the dissipation of a rain cell one hour later from the radar input. There are many feature attribution methods available; even a list of some of the more prevalent methods (*Saliency Maps* [37], *Integrated Gradients* [39], *GuidedGrad-CAM* [35] and *Layer-Wise Relevance Propagation(LRP)* [2] to name a few) can be extensive. This study selects the attribution method by quantitatively evaluating the completeness of the generated attributions following the incremental deletion criterion [27, 33]: the predictive performance of the model should decrease as the inputs are removed sequentially based on their importance, with the speed of decline faster at the initial stages of removal compared to the latter stages. After selecting a method, sample cases are analyzed by domain experts to evaluate user opinions on the generated results.

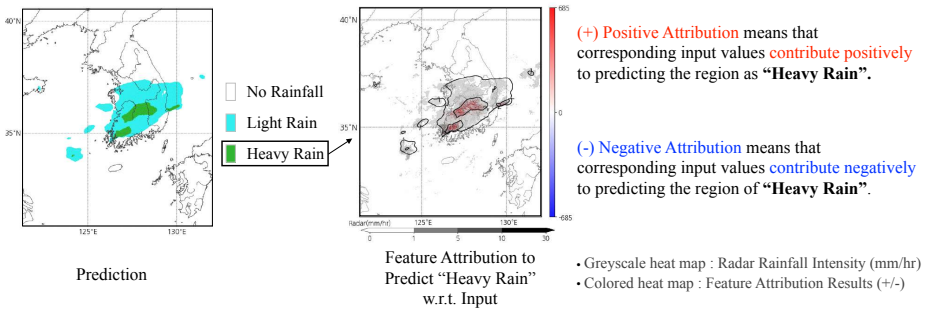

Fig. 7. Feature attribution. The heatmap describes the location and the degree of relevance of the inputs as the cause of the trained model prediction.

Quantitative Evaluation with Incremental Deletion. To quantitatively compare how well the feature importance maps from different methods reflect the true relative contributions of the features to the model predictions, the level of performance reduction is evaluated after eliminating the Top K% region of the input in the order of attribution value. A steeper decrease in performance implies greater fidelity. As shown in Figure 8, the integrated gradient method outperforms the other methods.

Qualitative Evaluation of Selected Attribution Method. To qualitatively evaluate the explanatory results, case-based anecdotal evidence has been analyzed through three consultations with external experts. Specifically, extreme precipitation cases are selected from the 2020 SOM-based classification study on the JJAS(June, July, August, and September which represent the period of the southwest monsoon) period in Korea by NIMS to match recognizable physical dynamics with attribution patterns.

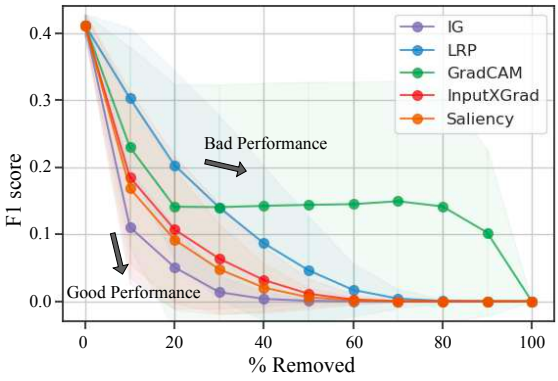

Fig. 8. Quantitative evaluation of the output reasoning explanation from different feature attribution methods.

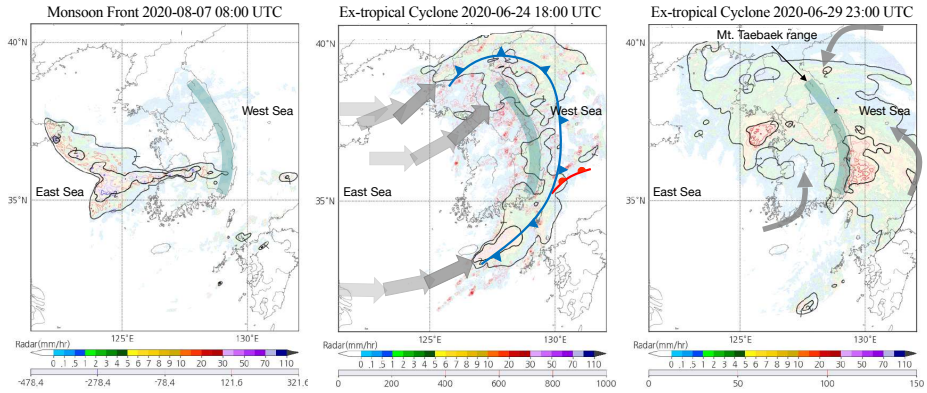

Fig. 9. Anecdotal evidence based on domain expert's case analysis of monsoon front and extratropical cyclones.

The leftmost case in Figure 9 is an extratropical cyclone system. The attribution map seems to describe the disappearance signal of fragmented convection cells moving in the direction opposite to the progression of the cold front(blue line). The middle figure is a case of the monsoon front, with a convection system moving from west to east. The attribution values are high at the edge of the radar area, most likely because the convective system is moving in from outside the effective range of the radar. This explanation can be considered an artifact. The rightmost case is an ex-tropical cyclone system. Moist and warm air from the East Sea and the West Sea blow inland, causing friction and rising along the Taebaek Mountain Range to result in convergence. The attribution heatmap seems to concur with this phenomenon, highlighting the corresponding area.

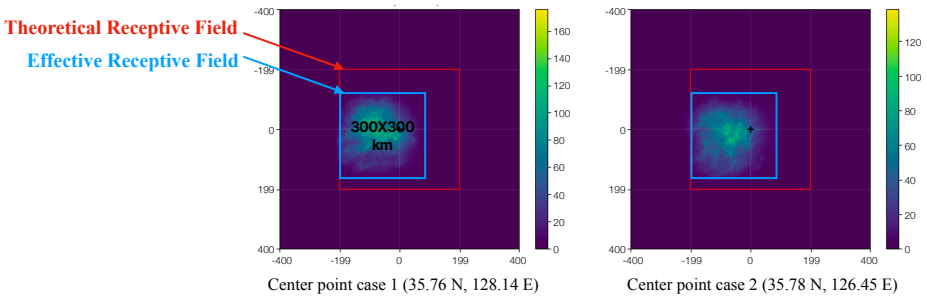

Fig. 10. Theoretical receptive field and effective receptive field of the target model. Due to the CNN structure, the maximum range of input region seen by a single output pixel is theoretically 398×398 km(approximately 200 km in radius). Depending on the learned parameters, the actual range is about 300×300 km(approximately 150 km in radius).

As an additional test, the receptive fields of the target model are identified using feature attribution. *Smooth Integrated Gradient* is applied on 75 samples and the average attribution map is used for evaluating the receptive field. As shown in Figure 10, the effective receptive field seems to be west-biased, which aligns with the fact that the westerlies are prevalent in Korea. The effective receptive field also has a radius of about 150 km. Assuming the maximum wind speed of 60 km per hour(about 16 m/s), the model may be making guesses when making predictions for three hours or later.

3.4 Explanation 3: Confidence Calibration

Confidence refers to the degree of certainty that a model has in its predictions. The certainty can be represented as a probability, and a well-calibrated model should be capable of assigning accurate confidence probabilities to its predictions. Unfortunately, deep learning models trained on negative log-likelihood (NLL) tend to exhibit overconfidence since it makes low-entropy distributions of the predictive classes [9] as demonstrated in Figure 11. In operational forecasting, a classification or segmentation model not only must be accurate but also indicate the point at which it is likely to be erred [14]. Probability calibration, the process of ensuring that the predicted probabilities of a model accurately reflect the true probabilities of the outcomes, can address this issue. [13]

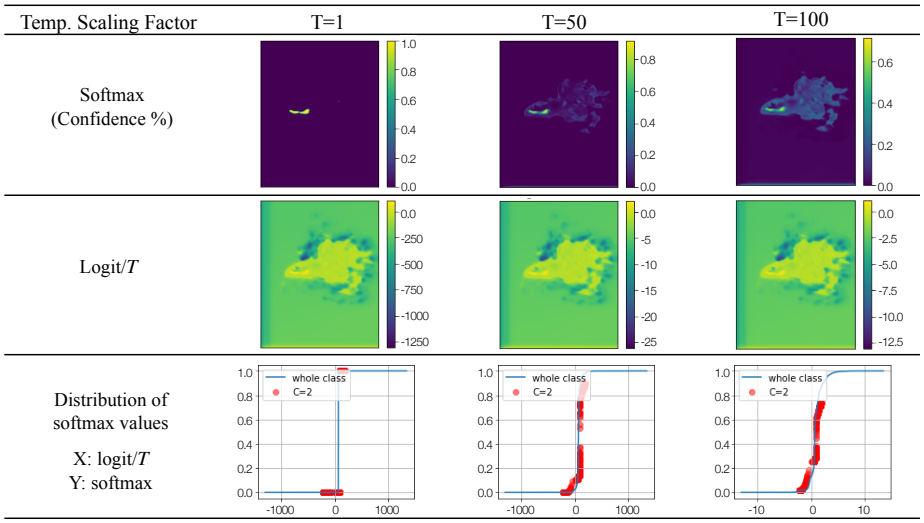

Fig. 11. The principle of temperature scaling. The softmax probability is scaled by a scalar parameter to reduce overconfidence by scaling the extreme logit values which occur near 0 or 100% of overconfidence. From left to right, the probability calibration progresses.

Table 1. Expected calibration error(ECE) of calibrated confidence on each lead time. The ECE is improved after calibration.

Lead Time	ECE	
	Before	After
1 hour	0.029	**0.010**
2 hour	0.099	**0.055**
3 hour	0.170	**0.037**
4 hour	0.232	**0.168**
5 hour	0.290	**0.109**
6 hour	0.320	**0.003**

Probability Calibration Methods. Probability calibration methods adjust the softmax of model logits as pseudo-probabilities. This paper uses the post-processing-based probability calibration methods which do not require re-training, making it suitable for quickly adjusting large-scale weather forecasting models. One of the simplest non-parametric approaches is *histogram binning*: all uncalibrated predictions are divided into mutually exclusive bins, enabling the selection of predictions that minimize bin-wise squared loss [40]. *Platt scaling* is a parametric calibration approach that employs the non-probabilistic predictions of a classifier as features for a logistic regression model [28]. *Temperature scaling(TS)* is a variant of Platt Scaling that employs a single scalar parameter $T > 0$ for all classes [13]. With the logit value z_i in each i-th pixel, the calibrated confidence is obtained as $\hat{q}_i(x, T) = \max_{k \in K} \sigma_{SM}(z_i/T)^{(k)}$.

Where k is the label index in K classes and σ_{SM} is softmax operation. The only learnable parameter T is optimized by the NLL. Since the maximum value of the softmax function σ_{SM} remains unaffected by T, the class prediction also remains unchanged. This consistency of model performance makes temperature scaling suitable for the task of probability calibration.

Local temperature scaling(LTS) [9] expands on the concept of TS in semantic segmentation tasks by introducing learnable parameters for individual image pixels. Their approach considers spatially varying temperature values and pixel-level changes. To achieve this, a mapping function is essential to train which takes logits $z(x)$ and the corresponding image sample x as inputs and generates scaling factors $T_i(x)$. These scaling factors are then divided by the logits $z_i(x)$. Formally,

$$\hat{q}_i(x, T_i(x)) = \max_{k \in K} \sigma_{SM}(z_i(x)/T_i(x))^{(k)} \tag{4}$$

where $T_i(x) \in \mathbb{R}^+$ is sample(image) and pixel dependent. We train the mapping functions for each lead time separately and employ a CNN, following a similar approach as described in the original paper. The mapping functions are optimized by minimizing the negative log-likelihood with respect to the validation dataset.

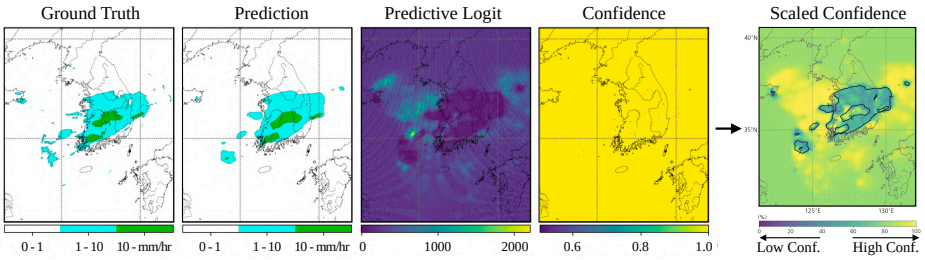

Fig. 12. The case of 2020-08-07 at 14:00 UTC with temperature scaling.

Evaluation for Probability Calibration. A commonly used measure of the probability calibration of a machine learning classifier is *expected calibration error*(ECE) [26], which estimates the difference between the predicted and the true probabilities. ECE is calculated by partitioning the range of predicted confidences into a set of bins and computing the weighted average difference between the average confidence $\mathrm{conf}(B_i)$ and the average accuracy $\mathrm{acc}(B_i)$ for each bin B_i as $\mathrm{ECE} = \sum_{b=1}^{B} \frac{n_b}{N} |\,\mathrm{acc}(B_i) - \mathrm{conf}(B_i)|$.

To utilize the ECE metric in the segmentation model, each pixel is considered as an individual sample as in [9]. To reduce computing costs, we randomly sample a predefined length of 250 ten times from a flattened array of confidence. Additionally, we masked ineffective areas in radar samples to improve the fidelity of the ECE metric by avoiding empty bins.

As shown in Table 1, the optimized LTS network improves the ECE scores after calibration for each of the six lead times, while maintaining the modified F1 scores. As demonstrated with an example in Figure 12, the LTS network diminishes the overconfidence in the predicted labels. The regions of heavy rain and no rain have high confidence scores rather than those of light rain while the predictive output seems to be similar to the ground truth.

3.5 Visualization: XAI Interface System

User interface design with XAI has been recently studied [5, 7]. In the design principles studied by Chromik et al. (2021) [7], XAI interfaces for users should provide progressive disclosure of explanatory information in order to avoid overwhelming users. This can be achieved through features such as tooltips or toggle buttons. Additionally, given that users are familiar with different explanation modalities such as natural language or visual explanations, users should be provided these modes to understand the information.

In this study, a pilot interface system has been established to display the explanations in a user-friendly manner, as shown in Figure 13 and 14. The explanation components consist of four parts:

Performance by Rainfall Type. After visualizing the input and prediction, the model performance explanation panel shows the test performance for the sample's rainfall type. A description of the training data is also provided.

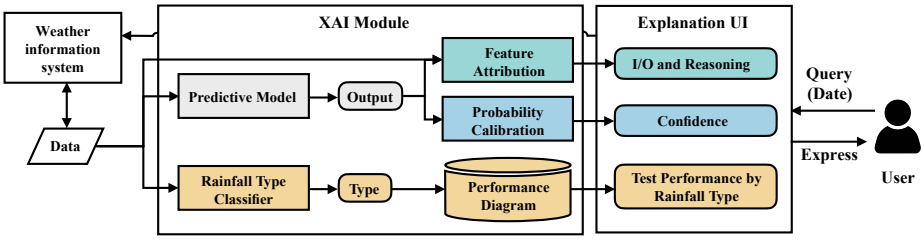

Fig. 13. Use case diagram for user interface and XAI modules.

Output Reasoning. The contribution of different target classes is computed simultaneously, allowing for comparison of the input contributions to no rain, light rain, and heavy rain classes.

Confidence. To explain confidence, a toggle key is provided that allows users to compare prediction and confidence results in individual grids.

Supplementary Materials. Based on user feedback, all results are presented along with other modalities that are excluded from the model inputs. The additional data allows for an increase in reliability as the users can verify their opinions on the generated results.

The text and color schemes in the visuals are expressed in plain language and domain terminology.

User Study on XAI Interface. Four forecasters in Korea Meteorological Agency participated in the user study. This user study aims to demonstrate the

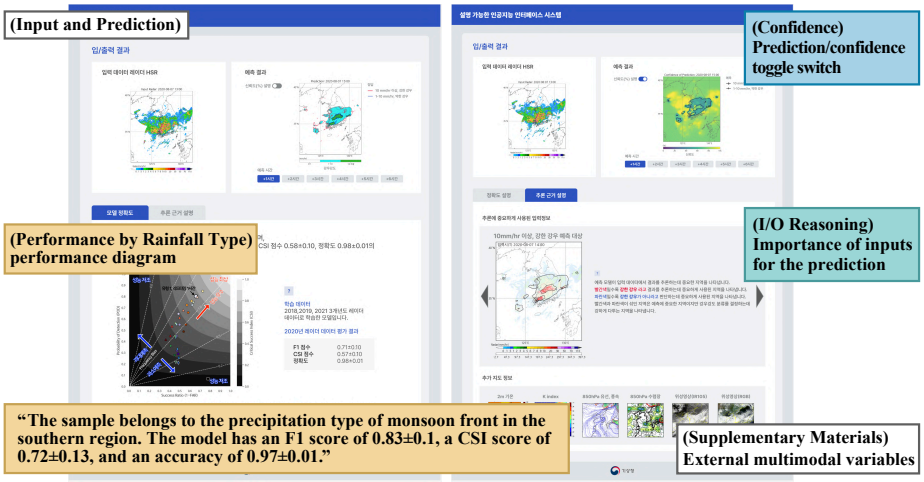

Fig. 14. Conceptual prototype of the interface system for the user-centered explanation. The demonstration is available(`https://figma.fun/LuhqIv`) in Korean

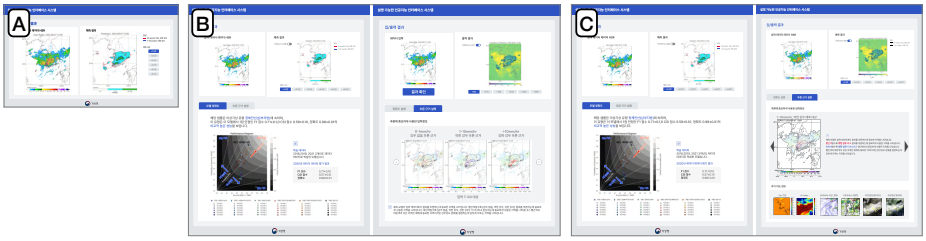

Fig. 15. Three prototypes for user survey. A demonstration is available for each: prototype A providing only prediction results (`https://figma.fun/uQcW9P`); prototype B adding three explanatory modules (`https://figma.fun/6n0CgH`); and prototype C including supplemented materials from user feedback and providing the simple and contracted information in the output reasoning module (`https://figma.fun/LuhqIv`).

interface system and elicit user feedback regarding their experience. The purpose of this survey is to qualitatively evaluate whether the explanatory modules, when provided alongside the predictions of an AI model, are useful to forecasters in practice. The survey assesses user experience based on three prototype interface systems (Q1-3) in Figure 15 and three types of explanatory modules (Q4-6). The participants answered 5-point Likert scale questions: Understandability *"Is the explanation easy to understand?"*, Usefulness *"Would use in practice?"*, Trustworthiness *"Can you trust the prediction?"*.

As results in Figure 16, compared to no explanation system A, the users experienced more trustful in the explanatory systems B and C which provide explanatory modules(B) and simplified explanation and additional thematic maps(C), respectively. The explanatory modules of the model performance by rainfall types (blue) and confidence (green) enhanced trustworthiness to some extent. Unfortunately, the users found the explanations to be difficult to understand in the output reasoning explanatory module (orange). The users also considered it unlikely to use output reasoning (orange) and confidence (green) explanatory modules in practice. The low usefulness of these explanatory modules was induced by the effort required to understand the information since forecasters often need to make decisions quickly. However, in order to increase user acceptance and usability in practice, the participants believed that it would be necessary to link the XAI interface system to the existing systems that forecasters use in the domestic meteorological agency. Also, forecasters found the research to be promising and were receptive to the idea of further investigating AI behavior. This response may provide a direction for future research, focusing on XAI receptiveness from the users.

4 Discussion

Through a series of meetings and interviews with the users, this study reduced the desired explanations into three main questions: model performance by rainfall type, inference reasoning, and output confidence. Based on the user needs

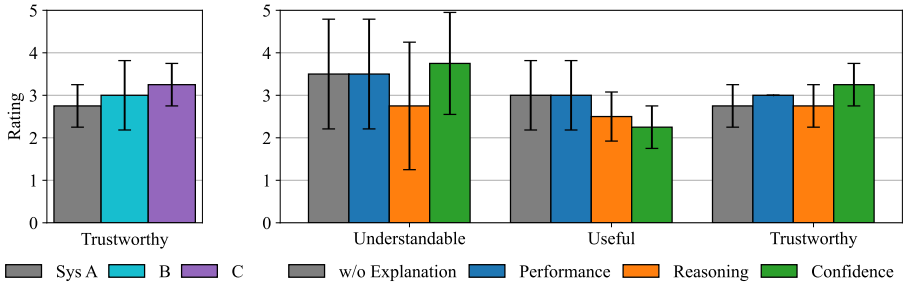

Fig. 16. Results from user experiment surveys. A comparison result of three prototypes of interface systems on trustworthiness(the left) and that of three explanatory modules on understandability, usefulness, and trustworthiness(the right).

and XAI algorithm mapping, a performance diagram with rainfall classifier, feature attribution, and probability calibration were selected as appropriate explanations for the requirements. Further analyses were performed to finalize the specific XAI methods in each category. Finally, three prototypes of the user interface were designed and feedback is received from the users. User experience survey results of the explanatory modules were promising on trustworthiness. Forecasters, however, requested high standards for actual use in practice since forecasters commonly need rapid decision-making.

One limitation of this study is that the overall process involved a specific set of users in Section 3.1; hence, the results may not cover the entirety of possible user requirements, creating a gap between XAI results and individual users' desired approaches as discussed in [21].

Another limitation is that for model performance by rainfall type in Section 3.2, the classifier shows limited performance due to a lack of samples with rainfall type labels. For actual implementation, it would be necessary to train the classifier with a larger dataset.

While the feature attribution methods in Section 3.3 can faithfully reflect model reasoning, even for distinct rainfall types, it can be challenging for experts to interpret. One reason for this difficulty is the model's reliance on uni-modal input features, restricting the feature attribution results to highlighting only the horizontal movement of convection cells. This issue may be addressed by using multi-modal data – in particular, since radar observations only represent the final outcomes of various physical mechanisms and the radar product used for training the target model provides only horizontal information, it would be ideal to include additional features that can provide this information.

In Section 3.5, users have provided feedback that the feature attribution explanations are hard to understand even if the explanations show high fidelity . This opinion indicates the need to measure the complexity of explanation results. Thus, user-centric XAI performance may need to reflect qualities of explanation besides faithfulness. Several previous works use Shannon entropy to measure

complexity in the image domain [4], but it is essential to recognize that the proxy variables in the weather domain have different characteristics due to their spatiotemporal context and may require different metrics of complexity.

Our pilot interface system is clickable, but it is a shallow-level user-interactive XAI(UXAI) system that becomes static after the completion of user-centric building procedures. Providing high-level interaction makes a potential area for future work to support explanations in response to feedback from the users such as interactive dialogue [7].

5 Conclusion

This study emphasizes the significance of involving users as key stakeholders in the design process of Explainable Artificial Intelligence (XAI) systems. Based on an analysis of user requirements in the meteorological domain and the mapping of these requirements to XAI methods, rainfall classification, feature attribution, and probability calibration are selected as suitable explanation. By presenting the model's performance for each rainfall type, users can judge the overall reliability of the corresponding AI model. Furthermore, sample cases of alignment with domain knowledge for feature attribution are identified. This investigation helps determine the practical applicability of feature attribution methods in meteorology. By providing confidence explanations for each output grid, users can assess the likelihood of output accuracy and decide the local reliability of individual predictions. Lastly, three prototypes of the user interface are designed and solicited feedback from users to ascertain the feasibility of integrating XAI into the forecasting system. This study may contribute to the literature as a use case of user-centered expression research.

References

1. Adadi, A., Berrada, M.: Peeking inside the black-box: a survey on explainable artificial intelligence (xai). IEEE access **6**, 52138–52160 (2018)
2. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one **10**(7), e0130140 (2015)
3. Başağaoğlu, H., Chakraborty, D., Lago, C.D., Gutierrez, L., Şahinli, M.A., Giacomoni, M., Furl, C., Mirchi, A., Moriasi, D., Şengör, S.S.: A review on interpretable and explainable artificial intelligence in hydroclimatic applications. Water **14**(8), 1230 (2022)
4. Bhatt, U., Weller, A., Moura, J.M.: Evaluating and aggregating feature-based model explanations. arXiv preprint arXiv:2005.00631 (2020)
5. Bradley, C., Wu, D., Tang, H., Singh, I., Wydant, K., Capps, B., Wong, K., Agostinelli, F., Irvin, M., Srivastava, B.: Explainable artificial intelligence (xai) user interface design for solving a rubik's cube. In: HCI International 2022–Late Breaking Posters: 24th International Conference on Human-Computer Interaction, HCII 2022, Virtual Event, June 26–July 1, 2022, Proceedings, Part II. pp. 605–612. Springer (2022)

6. Chaput, R., Cordier, A., Mille, A.: Explanation for humans, for machines, for human-machine interactions? In: AAAI-2021, Explainable Agency in Artificial Intelligence WS (2021)
7. Chromik, M., Butz, A.: Human-xai interaction: A review and design principles for explanation user interfaces. In: Human-Computer Interaction–INTERACT 2021: 18th IFIP TC 13 International Conference, Bari, Italy, August 30–September 3, 2021, Proceedings, Part II 18. pp. 619–640. Springer (2021)
8. Dell, M., Jones, B.F., Olken, B.A.: What do we learn from the weather? the new climate-economy literature. Journal of Economic literature **52**(3), 740–798 (2014)
9. Ding, Z., Han, X., Liu, P., Niethammer, M.: Local temperature scaling for probability calibration. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6889–6899 (2021)
10. Espeholt, L., Agrawal, S., Sønderby, C., Kumar, M., Heek, J., Bromberg, C., Gazen, C., Carver, R., Andrychowicz, M., Hickey, J., et al.: Deep learning for twelve hour precipitation forecasts. Nature communications **13**(1), 5145 (2022)
11. van der Geest, K., De Sherbinin, A., Kienberger, S., Zommers, Z., Sitati, A., Roberts, E., James, R.: The impacts of climate change on ecosystem services and resulting losses and damages to people and society. Loss and damage from climate change: Concepts, methods and policy options pp. 221–236 (2019)
12. Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L.: Explaining explanations: An overview of interpretability of machine learning. In: 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA). pp. 80–89. IEEE (2018)
13. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International Conference on Machine Learning. pp. 1321–1330 (2017)
14. Haynes, K., Lagerquist, R., McGraw, M., Musgrave, K., Ebert-Uphoff, I.: Creating and evaluating uncertainty estimates with neural networks for environmental-science applications. Artificial Intelligence for the Earth Systems pp. 1–58 (2023)
15. Kalnay, E.: Atmospheric modeling, data assimilation and predictability. Cambridge university press (2003)
16. Kim, C., Yun, S.Y.: Precipitation nowcasting using grid-based data in south korea region. In: 2020 International Conference on Data Mining Workshops (ICDMW). pp. 701–706. IEEE (2020)
17. Ko, J., Lee, K., Hwang, H., Oh, S.G., Son, S.W., Shin, K.: Effective training strategies for deep-learning-based precipitation nowcasting and estimation. Computers & Geosciences **161**, 105072 (2022)
18. Korea Meteorological Agency: Haneulsarang (2022), `https://www.kma.go.kr/download_01/kma_202002.pdf`
19. Liao, Q.V., Gruen, D., Miller, S.: Questioning the ai: informing design practices for explainable ai user experiences. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. pp. 1–15 (2020)
20. Liao, Q.V., Pribić, M., Han, J., Miller, S., Sow, D.: Question-driven design process for explainable ai user experiences. arXiv preprint arXiv:2104.03483 (2021)
21. Liao, Q.V., Varshney, K.R.: Human-centered explainable ai (xai): From algorithms to user experiences. arXiv preprint arXiv:2110.10790 (2021)
22. McGovern, A., Ebert-Uphoff, I., Gagne, D.J., Bostrom, A.: Why we need to focus on developing ethical, responsible, and trustworthy artificial intelligence approaches for environmental science. Environmental Data Science **1**, e6 (2022)
23. McGovern, A., Gagne, D.J., Williams, J.K., Brown, R.A., Basara, J.B.: Enhancing understanding and improving prediction of severe weather through spatiotemporal relational learning. Machine learning **95**, 27–50 (2014)

24. Mizutori, M., Guha-Sapir, D.: Economic losses, poverty and disasters 1998–2017. United Nations office for disaster risk reduction **4**, 9–15 (2017)
25. Murdoch, W.J., Singh, C., Kumbier, K., Abbasi-Asl, R., Yu, B.: Definitions, methods, and applications in interpretable machine learning. Proceedings of the National Academy of Sciences **116**(44), 22071–22080 (2019)
26. Naeini, M.P., Cooper, G.F., Hauskrecht, M.: Obtaining well calibrated probabilities using bayesian binning. Proceedings of the ... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence **2015**, 2901–2907 (2015)
27. Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., Seifert, C.: From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. arXiv preprint arXiv:2201.08164 (2022)
28. Platt, J.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers **10**(3), 61–74 (1999)
29. Rasp, S., Thuerey, N.: Data-driven medium-range weather prediction with a resnet pretrained on climate simulations: A new model for weatherbench. Journal of Advances in Modeling Earth Systems **13**(2), e2020MS002405 (2021)
30. Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., Fitzsimons, M., Athanassiadou, M., Kashem, S., Madge, S., et al.: Skilful precipitation nowcasting using deep generative models of radar. Nature **597**(7878), 672–677 (2021)
31. Ren, X., Li, X., Ren, K., Song, J., Xu, Z., Deng, K., Wang, X.: Deep learning-based weather prediction: a survey. Big Data Research **23**, 100178 (2021)
32. Roebber, P.J.: Visualizing multiple measures of forecast quality. Weather and Forecasting **24**(2), 601–608 (2009)
33. Samek, W., Binder, A., Montavon, G., Lapuschkin, S., Müller, K.R.: Evaluating the visualization of what a deep neural network has learned. IEEE transactions on neural networks and learning systems **28**(11), 2660–2673 (2016)
34. Schwalbe, G., Finzel, B.: A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. Data Mining and Knowledge Discovery pp. 1–59 (2023)
35. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
36. Shin, Y., Kim, J.H., Chun, H.Y., Jang, W., Son, S.W.: Classification of synoptic patterns with mesoscale mechanisms for downslope windstorms in korea using a self-organizing map. Journal of Geophysical Research: Atmospheres **127**(6), e2021JD035867 (2022)
37. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013)
38. Sønderby, C.K., Espeholt, L., Heek, J., Dehghani, M., Oliver, A., Salimans, T., Agrawal, S., Hickey, J., Kalchbrenner, N.: Metnet: A neural weather model for precipitation forecasting. arXiv preprint arXiv:2003.12140 (2020)
39. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: International conference on machine learning. pp. 3319–3328. PMLR (2017)
40. Zadrozny, B., Elkan, C.P.: Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In: International Conference on Machine Learning (2001)

41. Zhongming, Z., Linong, L., Xiaona, Y., Wangqiang, Z., Wei, L., et al.: Atlas of mortality and economic losses from weather, climate and water extremes (1970-2019). Weather Climate Water Temps Climate EAU (2021)