Evidential Deep Learning for Probabilistic Modelling of Extreme Storm Events

Ayush Khot[1,†], **Xihaier Luo**[2], **Ai Kagawa**[2], **Shinjae Yoo**[2]
[1]University of Illinois at Urbana-Champaign, [2]Brookhaven National Laboratory
{akhot2}@illinois.edu, {xluo, aik, sjyoo}@bnl.gov

Abstract

Uncertainty quantification (UQ) methods play an important role in reducing errors in weather forecasting. Conventional approaches in UQ for weather forecasting rely on generating an ensemble of forecasts from physics-based simulations to estimate the uncertainty. However, it is computationally expensive to generate many forecasts to predict real-time extreme weather events. Evidential Deep Learning (EDL) is an uncertainty-aware deep learning approach designed to provide confidence about its predictions using only one forecast. It treats learning as an evidence acquisition process where more evidence is interpreted as increased predictive confidence. We apply EDL to storm forecasting using real-world weather datasets and compare its performance with traditional methods. Our findings indicate that EDL not only reduces computational overhead but also enhances predictive uncertainty. This method opens up novel opportunities in research areas such as climate risk assessment, where quantifying the uncertainty about future climate is crucial. (Github: https://github.com/SULI24/edl-stormcast/)

1 Introduction

The study of climate and weather holds paramount importance due to its direct impact on natural ecosystems and human societies. Accurate weather predictions can significantly aid in disaster preparedness, agricultural planning, and energy management. Traditionally, numerical weather prediction (NWP) models have been the cornerstone of forecasting. These models simulate the atmosphere and its dynamics based on physical laws. These models, including the state-of-the-art High Resolution Ensemble Forecast (HREF) rainfall nowcasting model used in National Oceanic and Atmospheric Administration (NOAA) [37], rely on meticulous numerical simulation of physical models. These simulation-based systems fall short in the ability to incorporate signals from newly emerging geophysical observation systems [21], or take advantage of the Petabytes-scale Earth observation data [46]. In addition, they come with high computational costs and inherent complexities which can limit their scalability and accessibility [12, 27, 51].

In recent years, data-driven approaches, particularly deep learning, have emerged as powerful alternatives to traditional NWP models [38]. State-of-the-art methods in this domain have demonstrated impressive predictive capabilities by leveraging large datasets of historical weather patterns. Yet, a significant limitation of most existing deep learning approaches is their deterministic nature. Unlike traditional methods, these models typically do not provide measures of uncertainty in their predictions, which is crucial for risk management and decision-making in meteorology [18, 10].

Ensemble and Bayesian models address this gap by offering a framework to quantify prediction uncertainty, being especially used in weather forecasting problems [19, 30, 32]. Ensemble methods, for instance, generate multiple forecasts to capture a range of possible outcomes, enhancing the reliability of predictions [25]. Bayesian approaches, similarly, provide a probabilistic interpretation by considering the uncertainties in model parameters [15]. Nonetheless, both methods entail substantial

Machine Learning and the Physical Sciences Workshop, NeurIPS 2024.

computational overheads, primarily due to the need for multiple model evaluations or complex posterior calculations, making them less feasible for real-time forecasting applications. More information on related works are summarized in Appendix A.

To bridge this critical gap, we introduce an innovative approach utilizing evidential deep learning [40, 5]. This method extends the conventional deep learning framework to efficiently quantify uncertainties. Evidence deep learning, based on the concept of evidential theory, estimates the uncertainty directly during the learning process without the need for repetitive model runs or intricate probabilistic sampling. This allows for a significant reduction in computational demands while maintaining the ability to provide calibrated and interpretable uncertainty estimates alongside predictions.

Our empirical results validate that the proposed model not only achieves competitive accuracy in weather forecasting but also produces reliable and well-calibrated uncertainty measurements. These outcomes make a compelling case for the adoption of evidence deep learning in operational settings, where quick and dependable weather forecasts are crucial.

2 Methods

2.1 Problem Statement

The primary objective in weather forecasting is the accurate prediction of future atmospheric states based on a series of observed historical data. Let x_t denote the atmospheric state at time t, encapsulating various meteorological variables such as precipitation ($\mathrm{kg\,m^{-2}}$). The task is to forecast the future k time step states $x_{t+1}, x_{t+2}, \ldots, x_{t+k}$ by leveraging the historical sequence of the previous n time steps $x_{t-n}, x_{t-n+1}, \ldots, x_t$. The predictive task can be formalized as a function mapping from the domain of past atmospheric data to the domain of future states:

$$f : \mathcal{X}^{n+1} \to \mathcal{X}^k \tag{1}$$

where $\mathcal{X}$ represents the set of atmospheric states, n indicates the number of historical time steps utilized, k specifies the number of future time steps to be forecasted, and f is the predictive function.

2.2 Evidential Deep Learning

The central objective of this research is to develop an approximation of the function f using a sophisticated deep learning model, denoted as $\hat{f}$. For this purpose, we have selected EarthFormer, a state-of-the-art architecture known for its effectiveness in handling complex spatio-temporal data, as our primary model. However, the inherent challenge with such deterministic models lies in their inability to effectively capture the uncertainty intrinsic to meteorological phenomena, which is crucial given the chaotic nature of weather systems. To address this limitation, we propose to enhance the EarthFormer model by extending its capabilities into the probabilistic domain through the integration of evidential deep learning (EDL). Due to page constraints, the comprehensive details of the EarthFormer are relegated to the Appendix B. This section concentrates on the implementation of the EDL framework.

We are given a dataset, $\mathcal{D}$ with N paired training examples, $\mathcal{D} = \{\boldsymbol{x}_i, y_i\}_{i=1}^N$. In the deep evidential regression framework, the targets, y_i, are parameterized by a Gaussian distribution with an unknown mean and variance (μ, σ^2) [5, 40]. The conjugate prior distribution of (μ, σ^2) is then set to the Normal-Inverse-Gamma (`NIG`) distribution [33]. Deep evidential regression alters the model to output the parameters, $\boldsymbol{m} = (\gamma, \upsilon, \alpha, \beta)$, of the higher-order, evidential `NIG` distribution where $\gamma \in \mathbb{R}$, $\upsilon > 0$, $\alpha > 1$, $\beta > 0$. . Since $\boldsymbol{m}$ is composed of 4 parameters, the model needs four output neurons for every target parameter. To enforce the constraints on $(\upsilon, \alpha, \beta)$, we use a `softplus` activation (and additional $+1$ added to α since $\alpha > 1$). A linear activation is used for $\gamma \in \mathbb{R}$. Then given a `NIG` distribution, we can then compute the prediction, aleatoric, and epistemic uncertainty as:

$$\underbrace{\mathbb{E}[\mu] = \gamma}_{\text{prediction}}, \qquad \underbrace{\mathbb{E}[\sigma^2] = \tfrac{\beta}{\alpha-1}}_{\text{aleatoric uncertainty}}, \qquad \underbrace{\mathrm{Var}[\mu] = \tfrac{\beta}{\upsilon(\alpha-1)}}_{\text{epistemic uncertainty}}. \tag{2}$$

To ensure the model learns these parameters, the optimal loss function for the EDL model is composed of two primary components: the negative log likelihood, $\mathcal{L}_i^{\mathrm{NLL}}$, and the evidence regularizer, $\mathcal{L}_i^{\mathrm{R}}$. The authors of Ref. [5] show that by using Bayesian probability theory, $\mathcal{L}_i^{\text{NLL}}$ becomes a scaled Student's t-distribution parameterized as:

$$\mathcal{L}_i^{\text{NLL}}(\boldsymbol{w}) = -\log \text{St}\left(y_i; \gamma, \frac{\beta(1+\upsilon)}{\upsilon\,\alpha}, 2\alpha\right). \tag{3}$$

where St $\left(y; \mu_{\text{St}}, \sigma_{\text{St}}^2, \upsilon_{St}\right)$ is the Student-t distribution evaluated at y with location μ_{St}, scale σ_{St}^2, and υ_{St} degrees of freedom. Instead of using the KL Divergence in $\mathcal{L}_i^{\text{R}}$ like in the classification setting, the authors formulate a novel evidence regularizer as:

$$\mathcal{L}_i^{\text{R}}(\boldsymbol{w}) = |y_i - \gamma| \cdot (2\upsilon + \alpha) \tag{4}$$

The total loss, $\mathcal{L}_i(\boldsymbol{w})$, is composed of the two loss terms for maximizing and regularizing evidence, scaled by a regularization coefficient, λ,

$$\mathcal{L}_i(\boldsymbol{w}) = \mathcal{L}_i^{\text{NLL}}(\boldsymbol{w}) + \lambda\,\mathcal{L}_i^{\text{R}}(\boldsymbol{w}). \tag{5}$$

where $\lambda > 0$ is a regularization coefficient. Here, λ trades off uncertainty inflation with model fit. Setting $\lambda = 0$ yields an over-confident estimate while setting λ too high results in over-inflation. The authors of Ref. [40] proposed a dynamically scaled choice of λ to ensure a gradual increase of λ during the training process. This scaling allows the influence of the evidence regularizer to initially be limited, avoiding overly harsh penalties that could lead to model convergence towards an under-confident distribution prematurely. Then, as the model converges, the evidence regularizer becomes more prominent, guiding the model towards a more accurate uncertainty quantification.

3 Experiments

3.1 Experimental Setup

Data We utilize the Storm EVent ImageRy (SEVIR) dataset [46]. Specifically, the task is defined as precipitation nowcasting by predicting the next 12 images, each representing a 5-minute interval, in the sequence given 13 images, or 65 minutes, additionally detailed in Appendix C.1.

Baselines The proposed EDL is assessed against two state-of-the-art models: ensemble methods [25] and Monte Carlo (MC) Dropout [15]. Both methods require multiple inferences to estimate the uncertainty. As a result, we use 10 different inference passes. Details are available in Appendix C.2.

3.2 Results

Model Accuracy All training is done on two NVIDIA H100 GPUs, and all testing is done on 1 NVIDIA A100 GPU. We first calculate the Critical Success Index (CSI), a widely used metric in precipitation nowcasting that evaluates forecast accuracy by comparing correctly predicted events to the total number of predicted or observed events. CSI is usually defined as $\frac{\#\text{Hits}}{\#\text{Hits}+\#\text{Misses}+\#\text{FalseAlarms}}$, where #Hits (truth=1, pred=1), #Misses (truth=1, pred=0), and #FalseAlarms (truth=0, pred=1) are determined after rescaling predictions and ground truth to 0-255 and binarizing at thresholds $[16, 74, 133, 160, 181, 219]$. CSI ranges from 0 to 1, with 1 indicating a perfect forecast. In Table 1, we report CSI values across different thresholds. Results indicate that EDL performs relatively better at lower thresholds, though it lags behind Ensemble and MC Dropout methods. Notably, using initial weights pretrained without EDL significantly boosts EDL performance, especially at higher thresholds like CSI-181 and CSI-160.

Table 1: Prediction performance comparison. Note that P-EDL refers to EDL with pretrained weights optimized using MSE loss.

Model	Metrics					
	CSI-219 ↑	CSI-181 ↑	CSI-160 ↑	CSI-133 ↑	CSI-74 ↑	CSI-16 ↑
Ensemble	0.1436	0.2613	0.3081	0.4225	0.6947	0.7666
MC Dropout	0.1436	0.2613	0.3081	0.4225	0.6947	0.7666
EDL	0.0027	0.0726	0.1530	0.3303	0.6676	0.7358
P-EDL	0.0066	0.1215	0.2205	0.3682	0.6416	0.7562

We also compute the mean squared error (MSE) as a function of prediction lead time. As shown in Figure 1, the MSE for all models increases with longer lead times, highlighting the challenge of maintaining accuracy over extended predictions. This trend also suggests that the Ensemble and MC Dropout methods may be preferable when the primary focus is on minimizing prediction error over time.

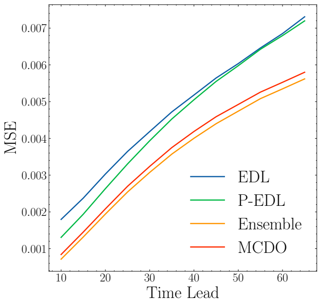

Figure 1: Plot of average MSE for varying forecasts time leads.

Model Efficiency We further assess the computational efficiency of each model by measuring average inference time and GFLOPS. As shown in Figure 2, EDL consistently outperforms both MC Dropout and Ensemble in terms of GFLOPS and inference time. This advantage arises because EDL conducts uncertainty analysis directly within a single model, whereas MC Dropout and Ensemble require multiple models or particles, necessitating several inferences to estimate uncertainty. Additionally, EDL and MC Dropout have a comparable number of parameters, both significantly fewer than Ensemble, which relies on multiple models. However, MC Dropout's need for repeated runs to generate uncertainty estimates substantially increases its computational cost.

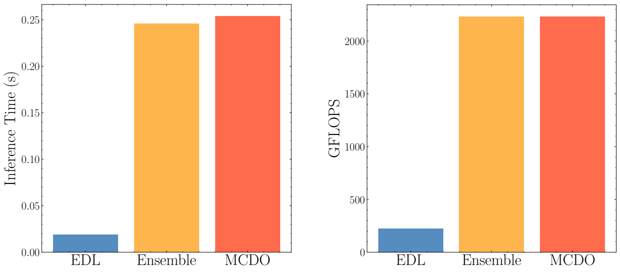

Figure 2: Histograms of the inference time and GFLOPS.

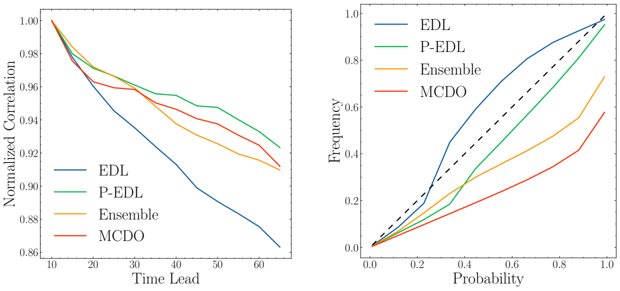

Figure 3: Uncertainty analysis results: The left panel displays the normalized correlation between uncertainty and MSE, while the right panel shows the reliability curves. Models closer to the dashed $y = x$ line exhibit well-calibrated uncertainty.

Uncertainty Diagnosis In our final analysis, we assess the quality of the estimated uncertainty across different models. To begin, we investigate the relationship between uncertainty and model accuracy by plotting the normalized correlation between uncertainty and MSE at varying forecast time leads (Figure 3). As expected, the correlation generally decreases over time, reflecting the growing difficulty in making accurate predictions as the forecast horizon extends. Notably, P-EDL maintains a higher normalized correlation over time, suggesting more stable uncertainty estimates, whereas EDL experiences a sharp decline, indicating a potential loss in calibration with increasing lead time. We also evaluate calibration using a reliability diagram, where a perfectly calibrated model aligns with the dashed $y = x$ line. EDL demonstrates strong calibration, with P-EDL slightly less accurate. By utilizing pretrained weights, P-EDL prioritizes the accuracy of the prediction rather than the accuracy of the uncertainty. As a result, the accuracy is higher but it has worse uncertainty calibration. In contrast, Ensemble and MC Dropout deviate further from the optimal line, highlighting the superior uncertainty estimation of EDL. Additional uncertainty plots are provided in the Appendix D.

4 Conclusion

We propose the use of EDL for storm forecasting as an effective measure of model uncertainty. Although the model has a slightly worse MSE and `CSI` values, the use of only one model and one forecast makes it much less time-consuming and computationally expensive than both Ensemble and MC Dropout. The uncertainty from EDL has also proven to be more well-calibrated than traditional methods. We propose the use of pretrained weights on MSE loss for EDL, P-EDL, to ensure that the MSE loss significantly decreases and it maintains higher normalized correlation with increased leading time. P-EDL ensures similarly well-calibrated uncertainty. Uncertainty calibration and computational is very important when dealing with reliable, real-time predictions, but it shouldn't come at the expense of accuracy. We will work on subsequent methods to maintain accuracy in EDL. In the future, EDL could be incorporated to refine the uncertainty estimates across ensemble members, providing a more robust assessment of model predictions by weighting them based on their evidential support. In addition, other data distributions, such as Poisson or a Gamma distribution, could be adopted within EDL.

Acknowledgments and Disclosure of Funding

This work was supported in part by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Science Undergraduate Laboratory Internships (SULI) program. This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 using NERSC award ASCR-ERCAP0027539. Xihaier Luo was supported by funding from the Advanced Scientific Computing Research program within the Department of Energy's Office of Science, under project No. DE-SC0012704. SCIDAC IMPACTS FWP # CC127. Ayush Khot and Ai Kagawa were additionally supported by the concept development funding from the Computing and Data Sciences Directorate at Brookhaven National Laboratory. The authors acknowledge Nathan Urban for useful discussions, comments and suggestions.

References

[1] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, 2021.

[2] Abdullah A. Abdullah, Masoud M. Hassan, and Yaseen T. Mustafa. A review on bayesian deep learning in healthcare: Applications and challenges. *IEEE Access*, 10:36538–36562, 2022.

[3] Abdullah A. Abdullah, Masoud M. Hassan, and Yaseen T. Mustafa. Leveraging bayesian deep learning and ensemble methods for uncertainty quantification in image classification: A ranking-based approach. *Heliyon*, 10(2):e24188, 2024.

[4] Daniel Althoff, Lineu Neiva Rodrigues, and Helizani Couto Bazame. Uncertainty quantification for hydrological models based on neural networks: the dropout ensemble. *Stoch. Environ. Res. Risk Assess.*, 35(5):1051–1067, May 2021.

[5] Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. *Advances in Neural Information Processing Systems*, 33, 2020.

[6] Amirmasoud Amini, Mehri Dolatshahi, and Reza Kerachian. Adaptive precipitation nowcasting using deep learning and ensemble modeling. *Journal of Hydrology*, 612:128197, 2022.

[7] Sojung An, Tae-Jin Oh, Eunha Sohn, and Donghyun Kim. Deep learning for precipitation nowcasting: A survey from the perspective of time series forecasting, 2024.

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

[9] Cong Bai, Feng Sun, Jinglin Zhang, Yi Song, and Shengyong Chen. Rainformer: Features extraction balanced network for radar-based precipitation nowcasting. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022.

[10] Christopher Bülte, Nina Horat, Julian Quinting, and Sebastian Lerch. Uncertainty quantification for data-driven weather models, 2024.

[11] Alp Kucukelbir David M. Blei and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

[12] ECMWF. *IFS Documentation CY47R3 - Part V Ensemble prediction system*. Number 5. ECMWF, 09/2021 2021.

[13] Lasse Espeholt, Shreya Agrawal, Casper Sønderby, Manoj Kumar, Jonathan Heek, Carla Bromberg, Cenk Gazen, Jason Hickey, Aaron Bell, and Nal Kalchbrenner. Skillful twelve hour precipitation forecasts using large context neural networks. *arXiv preprint arXiv:2111.07470*, 2021.

[14] Jesús García Fernández and Siamak Mehrkanoon. Broad-unet: Multi-scale feature learning for nowcasting tasks. *Neural Networks*, 144:419–427, 2021.

[15] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 1050–1059. JMLR.org, 2016.

[16] Zhihan Gao, Xingjian Shi, Boran Han, Hao Wang, Xiaoyong Jin, Danielle C Maddix, Yi Zhu, Mu Li, and Bernie Wang. Prediff: Precipitation nowcasting with latent diffusion models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[17] Zhihan Gao, Xingjian Shi, Hao Wang, Yi Zhu, Yuyang Wang, Mu Li, and Dit-Yan Yeung. Earthformer: Exploring space-time transformers for earth system forecasting. In *NeurIPS*, 2022.

[18] Tilmann Gneiting and Matthias Katzfuss. Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1(Volume 1, 2014):125–151, 2014.

[19] Tilmann Gneiting and Adrian E. Raftery. Weather forecasting with ensemble methods. *Science*, 310(5746):248–249, 2005.

[20] Junchao Gong, Lei Bai, Peng Ye, Wanghan Xu, Na Liu, Jianhua Dai, Xiaokang Yang, and Wanli Ouyang. Cascast: Skillful high-resolution precipitation nowcasting via cascaded modelling. *arXiv preprint arXiv:2402.04290*, 2024.

[21] Steven J Goodman, Timothy J Schmit, Jaime Daniels, and Robert J Redmon. *The GOES-R series: a new generation of geostationary environmental satellites*. Elsevier, 2019.

[22] Vincent Le Guen and Nicolas Thome. Disentangling physical dynamics from unknown factors for unsupervised video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11474–11484, 2020.

[23] John Guibas, Morteza Mardani, Zongyi Li, Andrew Tao, Anima Anandkumar, and Bryan Catanzaro. Adaptive fourier neural operators: Efficient token mixers for transformers. *arXiv preprint arXiv:2111.13587*, 2021.

[24] Fredrik K. Gustafsson, Martin Danelljan, and Thomas B. Schön. Evaluating scalable bayesian deep learning methods for robust computer vision, 2020.

[25] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6405–6416, Red Hook, NY, USA, 2017. Curran Associates Inc.

[26] Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, et al. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023.

[27] Martin Leutbecher. Ensemble size: How suboptimal is less than infinity? *Q. J. R. Meteorol. Soc.*, 145:107–128, 9 2019. https://doi.org/10.1002/qj.3387.

[28] Hao Li and Jianan Liu. 3d high-quality magnetic resonance image restoration in clinics using deep learning. *ArXiv*, abs/2111.14259, 2021.

[29] 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*, pages 10012–10022, 2021.

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

[31] Patrick L. McDermott and Christopher K. Wikle. Deep echo state networks with uncertainty quantification for spatio-temporal forecasting. *Environmetrics*, 30(3):e2553, 2019. e2553 env.2553.

[32] Prabha Shreeraj Nair and G. Ezhilarasan. Bayesian models for weather prediction: Using remote sensing data to improve forecast accuracy. In Om Prakash Verma, Lipo Wang, Rajesh Kumar, and Anupam Yadav, editors, *Machine Intelligence for Research and Innovations*, pages 327–343, Singapore, 2024. Springer Nature Singapore.

[33] Giorgio Parisi. *Statistical field theory*. Addison-Wesley, 1988.

[34] Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, et al. FourCastNet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*, 2022.

[35] Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R. Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, Remi Lam, and Matthew Willson. Gencast: Diffusion-based ensemble forecasting for medium-range weather, 2024.

[36] Adrian E. Raftery, Tilmann Gneiting, Fadoua Balabdaoui, and Michael Polakowski. Using bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, 133(5):1155 – 1174, 2005.

[37] Suman Ravuri, Karel Lenc, Matthew Willson, Dmitry Kangin, Remi Lam, Piotr Mirowski, Megan Fitzsimons, Maria Athanassiadou, Sheleem Kashem, Sam Madge, et al. Skilful precipitation nowcasting using deep generative models of radar. *Nature*, 597(7878):672–677, 2021.

[38] Markus Reichstein, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais, et al. Deep learning and process understanding for data-driven earth system science. *Nature*, 566(7743):195–204, 2019.

[39] John S. Schreck, David John Gagne II au2, Charlie Becker, William E. Chapman, Kim Elmore, Da Fan, Gabrielle Gantos, Eliot Kim, Dhamma Kimpara, Thomas Martin, Maria J. Molina, Vanessa M. Pryzbylo, Jacob Radford, Belen Saavedra, Justin Willson, and Christopher Wirz. Evidential deep learning: Enhancing predictive uncertainty estimation for earth system science applications, 2024.

[40] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In *Advances in Neural Information Processing Systems*, pages 3179–3189, 2018.

[41] 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. In *NeurIPS*, volume 28, 2015.

[42] J. Mc Lean Sloughter, Adrian E. Raftery, Tilmann Gneiting, and Chris Fraley. Probabilistic quantitative precipitation forecasting using bayesian model averaging. *Monthly Weather Review*, 135(9):3209 – 3220, 2007.

[43] Bofan Song, Sumsum Sunny, Shaobai Li, G Keerthi, Sanjana Patrick, Nirza Mukhia, Shubha Gurudath, Subhashini Raghavan, Pramila Mendonca, Tsusennaro, Shirley T Leivon, Trupti Kolur, Vivek Shetty, Vidya Bushan R, Rohan Ramesh, Vijay Pillai, Alben Sigamani, Amritha

Suresh, moni Abraham Kuriakose, Praveen Birur, and Rongguang Liang. Reliable oral cancer classification framework with bayesian deep learning. In *Frontiers in Optics / Laser Science*, page JM6B.18. Optica Publishing Group, 2020.

[44] Maria Antonia Sunyer, Henrik Madsen, Dan Rosbjerg, and Karsten Arnbjerg-Nielsen. A bayesian approach for uncertainty quantification of extreme precipitation projections including climate model interdependency and nonstationary bias. *Journal of Climate*, 27(18):7113 – 7132, 2014.

[45] Dennis Ulmer. A survey on evidential deep learning for single-pass uncertainty estimation. *ArXiv*, abs/2110.03051, 2021.

[46] Mark Veillette, Siddharth Samsi, and Chris Mattioli. SEVIR: A storm event imagery dataset for deep learning applications in radar and satellite meteorology. *Advances in Neural Information Processing Systems*, 33:22009–22019, 2020.

[47] Yunbo Wang, Lu Jiang, Ming-Hsuan Yang, Li-Jia Li, Mingsheng Long, and Li Fei-Fei. Eidetic 3D LSTM: A model for video prediction and beyond. In *International conference on learning representations*, 2018.

[48] Yunbo Wang, Haixu Wu, Jianjin Zhang, Zhifeng Gao, Jianmin Wang, Philip Yu, and Mingsheng Long. PredRNN: A recurrent neural network for spatiotemporal predictive learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[49] Demin Yu, Xutao Li, Yunming Ye, Baoquan Zhang, Chuyao Luo, Kuai Dai, Rui Wang, and Xunlai Chen. Diffcast: A unified framework via residual diffusion for precipitation nowcasting. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[50] Yuchen Zhang, Mingsheng Long, Kaiyuan Chen, Lanxiang Xing, Ronghua Jin, Michael I. Jordan, and Jianmin Wang. Skilful nowcasting of extreme precipitation with nowcastnet. *Nature*, 619(7970):526–532, Jul 2023.

[51] Xiaqiong Zhou, Yuejian Zhu, Dingchen Hou, Bing Fu, Wei Li, Hong Guan, Eric Sinsky, Walter Kolczynski, Xianwu Xue, Yan Luo, Jiayi Peng, Bo Yang, Vijay Tallapragada, and Philip Pegion. The development of the NCEP global ensemble forecast system version 12. *Weather Forecast.*, 37(6):1069–1084, June 2022.

A Related Works

Deep Learning-based Weather Prediction. Traditional DL approaches to weather forecasting involve CNN and RNN. U-Net based architectures that include CNN have been applied to sea ice forecasting [8], precipitation nowcasting [46, 50], and cloud cover nowcasting [14]. Shi et al. [41] proposed ConvLSTM that leverages CNN and LSTM for precipitation nowcasting. Wang et al. [48] proposed PredRNN which extends the predictive capabilities of ConvLSTM by introducing a spatiotemporal memory architecture. To better learn long-term high-level relations, Wang et al. [47] proposed E3D-LSTM that integrates 3D CNN with LSTM. To disentangle PDE dynamics from unknown complementary information, Guen et al. proposed PhyDNet [22] which incorporates a new recurrent physical cell to perform PDE-constrained prediction in latent space. Espeholt et al. [13] proposed MetNet-2, based on ConvLSTM and dilated CNN, that outperforms HREF in precipitation forecasting. Very recently, there are works that have implemented Transformer for solving weather forecasting problems [7] Pathak et al. [34] proposed the FourCastNet for global weather forecasting, which is based on Adaptive Fourier Neural Operators (AFNO) [23]. Bai et al. [9] proposed Rainformer for precipitation nowcasting, which is based on an architecture that combines CNN and Swin-Transformer [29]. Later, Gao et al. proposed the Earthformer [17] which implements cuboid attention to reduce computional expense. Pangu-Weather leverages a 3D Vision Transformer that separates the input into cubes to predict in a medium time range scale ($5 - 10$ days). There are also other GNN models like GraphCast [26] and diffusion models like GenCast [35], but precipitation forecasting has not been their strong point. Diffusion-based models are probabilistic, but sequentially denoising the input over multiple steps can be computationally expensive [16, 49, 20].

Probabilistic Modeling It is preferred to extend deterministic models to incorporate uncertainty, generating a distribution of possible outcomes rather than a single outcome. There are a variety of popular techniques to incorporate uncertainty: **(1) Ensemble Methods:** By combining multiple models or fit one model with diverse hyperparameters, we can get better generalization performance in the final prediction [25]. Popular in NWP models [19], ensemble methods has gained traction in deep learning [1]. Some examples of its applications include computer vision [1], spatiotermporal forecasting [31], mechinal machinery [4], and precipitation nowcasting [6]. However, training multiple models can be time-consuming and computationally intensive. **(2) Bayesian Methods:** Instead of using multiple models, Bayesian deep learning (BDL) incorporates Bayesian concepts to quantify uncertainty effectively. BDL models utilize a posterior probability distribution that relies on prior knowledge distribution and the likelihood of the data being utilized [15]. This framework represents model weights as random variables [3], and the uncertainty can be generated by using the prior and likelihood to sample from the posterior distribution. The most popular techniques are Monte Carlo Dropout [15] and Variational Inference (VI) [11]. These techniques are popular in computer vision [43, 24], healthcare [2], and weather forecasting [30]. It is also very popular in precipitation nowcasting [42, 44, 36]. However, it is difficult to scale due to the multiple predictoins needed to approximate uncertainty. **(3) Evidential Deep Learning Methods:** Evidential Deep Learning (EDL) is a novel method to quantify uncertainty directly by modeling the evidence supporting different outcomes [5, 40]. By directly using one model and a single forward pass, EDL offers a comprehensive framework to learn the uncertainty. This technique has been used in healthcare [28], computer vision [45], and Earth system modeling [39]. Schreck et al. [39] also apply EDL in Earth system science applications, but they focus more on Earth system modeling rather than forecasting. Instead, we focus on weather forecasting problems. We examine the computational expense and the uncertainty calibration of EDL compared with baseline methods.

B Model Architecture

The Earthformer model employs a sequence of atmospheric state vectors as its input, representing the dynamic meteorological conditions over time:

$$\mathcal{X}^{n+1} = [x_{t-n}, \ldots, x_t] \tag{6}$$

Each vector x in the sequence encapsulates essential atmospheric variables such as temperature, pressure, and humidity, corresponding to specific timestamps.

Input Embedding: Each input vector x_t undergoes a transformation into a higher-dimensional feature space to facilitate more complex interactions and learning. This transformation is realized through an embedding layer defined by:

$$\text{Emb}(x_t) = W_e x_t + b_e \tag{7}$$

where W_e and b_e represent the embedding weight matrix and bias vector, respectively, both of which are parameters learned during training.

Positional Encoding: To incorporate the temporal sequence information essential for forecasting, positional encodings are added to the input embeddings. These encodings are computed using trigonometric functions:

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right), \quad PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right) \tag{8}$$

where pos indicates the position in the sequence and i denotes the dimension index. The embedded inputs are then modified as:

$$H_0 = \text{Emb}(X) + PE \tag{9}$$

Cuboid Attention: EarthFormer is a hierarchial Transformer encoder-decoder based on Cuboid Attention. Traditional Transformers compute attention scores using a query (Q), key (K), and value (V) framework, commonly expressed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{10}$$

In this equation, Q, K, and V are matrices derived from the input data, where d_k represents the dimensionality of the keys, facilitating the scaling of dot products.

Cuboid attention modifies this framework to address the complexities of 3D data, focusing on the spatial, depth, and temporal dimensions. It is computed by reshaping the embeded inputs H_0 to emphasize its three-dimensional structure:

$$Q, K, V = \text{reshape}(H_0), \quad \text{to shape } (D, H, W, d_k) \tag{11}$$

The attention scores are then calculated across blocks or cuboids within the data:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{\text{reshape}(Q) \cdot \text{reshape}(K)^T}{\sqrt{d_k}}\right) \cdot \text{reshape}(V) \tag{12}$$

This reshaping and attention computation enables the model to focus on interactions not just within each plane, but across all three dimensions. To synthesize the information from different dimensions, the attention outputs are combined:

$$H = \frac{1}{3}(\text{Attention}_D(Q, K, V) + \text{Attention}_H(Q, K, V) + \text{Attention}_W(Q, K, V)) \tag{13}$$

Where Attention_D, Attention_H, and Attention_W are the attention computations performed along the depth, height, and width dimensions, respectively.

Layer Normalization and Feed-Forward Network: After attention processing, the output undergoes layer normalization to stabilize the training process, followed by a feed-forward network:

$$H' = \text{LayerNorm}(H + \text{Attention}(H)), \quad H'' = \text{LayerNorm}(H' + \text{FFN}(H')) \tag{14}$$

The feed-forward network (FFN) consists of two linear transformations with a nonlinear activation function in between, enhancing the model's ability to capture non-linear relationships.

Output: The output from the final Transformer block H'' is decoded to predict future atmospheric states, which are essential for accurate weather forecasting:

$$\mathcal{X}^k = [x_{t+1}, \ldots, x_{t+k}] = \text{Decoder}(H'') \tag{15}$$

where $\mathcal{X}^k$ denotes the predicted future states, providing crucial insights into upcoming weather conditions.

C Additional Details on Experiments

C.1 Data

We utilize the Storm EVent ImageRy (SEVIR) dataset [46], a widely used benchmark for meterological applications. This spatiotemporally aligned dataset contains over 10,000 weather events, each consisting of 384 km × 384 km image sequences that span over 4 hours. Images in SEVIR were sampled and aligned across five different data types: three channels (C02, C09, C13) from the GOES-16 advanced baseline imager, NEXRAD vertically integrated liquid mosaics, and GOES-16 Geostationary Lightning Mapper (GLM) flashes. This dataset supports a variety of deep learning research in meteorological applications like precipitation nowcasting, synethic radar generatino, front detection, and more. We use SEVIR for precipitation nowcasting by predicting the next 12 images, or 60 minutes, in the sequence given 13 images, or 65 minutes. Like the authors of Ref. [17], we normalized the data to the range $[0, 1]$. Figure 4 gives an example of the SEVIR VIL frame sequences.

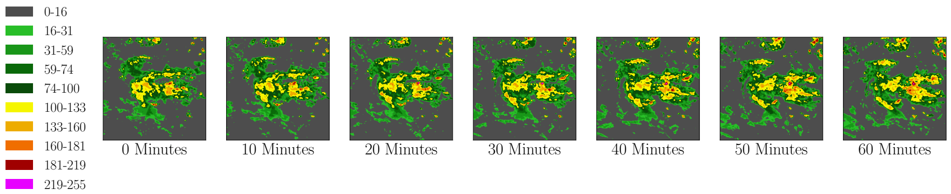

Figure 4: Example Vertically Integrated Liquid (VIL) observation sequence from the Storm EVent ImageRy (SEVIR) dataset. The observation intensity is mapped to pixel value of the range 0-255. The larger value indicates the higher precipitation intensity.

C.2 Baselines

Traditionally, ensemble methods [25] and Monte Carlo (MC) Dropout [15] have been popular techniques for estimating epistemic uncertainty in neural networks. Ensemble methods involve training multiple models independently and using the variance of the outputs to evaluate uncertainty. MC Dropout utilizes dropout layers both during training and inference to stimulate the effect of Bayesian inference, thus providing a stochastic basis for uncertainty estimation. Both methods are computationally intensive as they require multiple inferences to estimate the uncertainty, reflecting a significant trade-off between computational efficiency and accuracy. This makes them less prone to be used in real-time due to the large memory and computational expense. For our purposes, we use 10 different inference passes for both Ensemble and MC Dropout.

D Uncertainty Plots

We plot the uncertainties for both the EDL and P-EDL models below using test data. We display the target and output values as well as the Root Mean Square Error (RMSE) and epistemic uncertainty for visualization purposes. We visualize these values over forecasts times from 10 to 40 minutes in the future. The uncertainty tends to get less detailed as the forecast time increases, and the errors also tend to increase. Most of the uncertainties are near the edges of the storms, and they are often concentrated near high VIL values, which the model is insufficient at predicting.

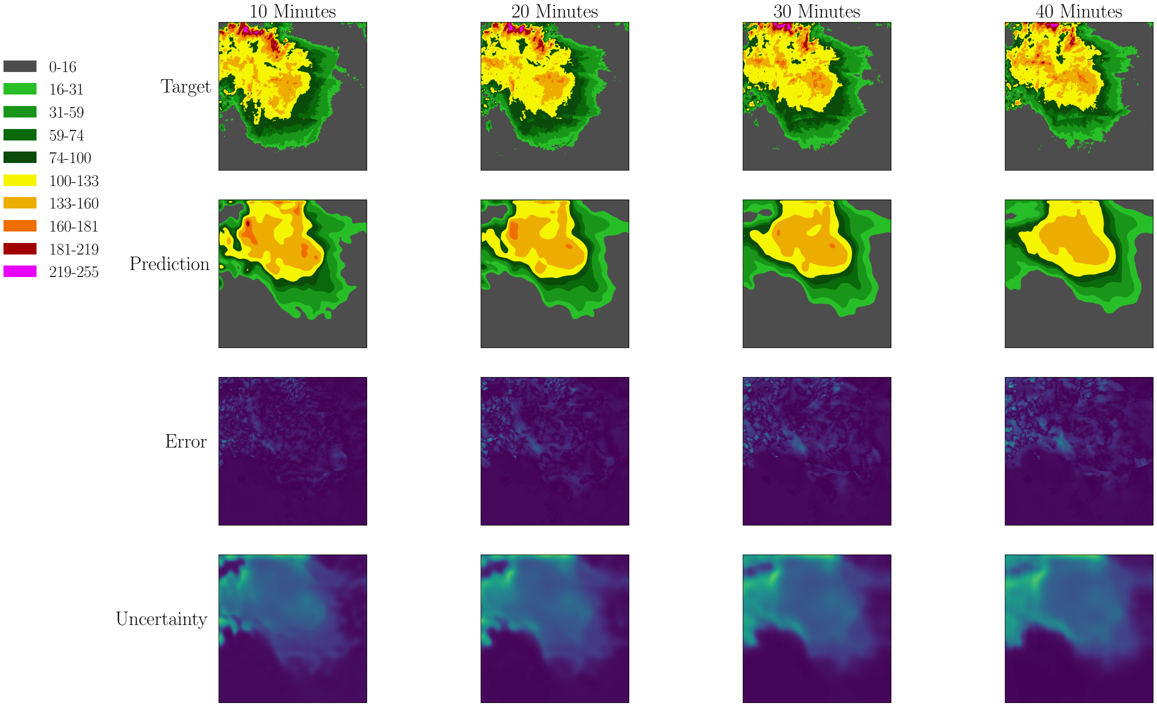

Figure 5: Plot of error and uncertainties for EDL model

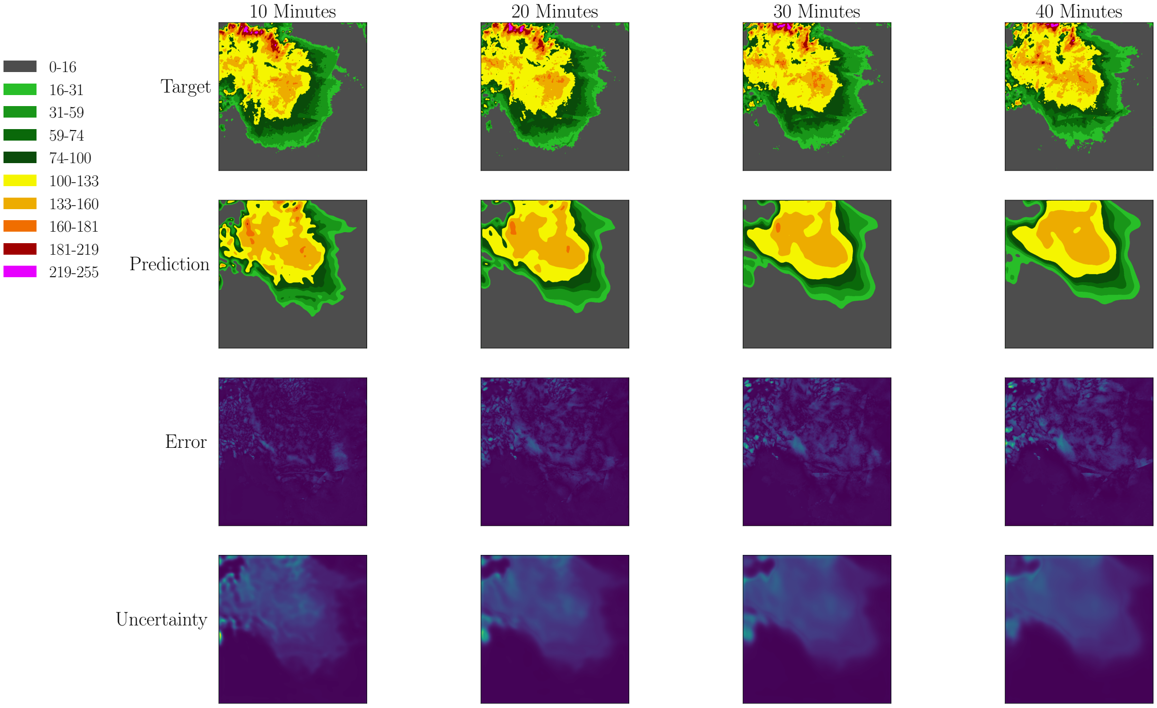

Figure 6: Plot of error and uncertainties for P-EDL model

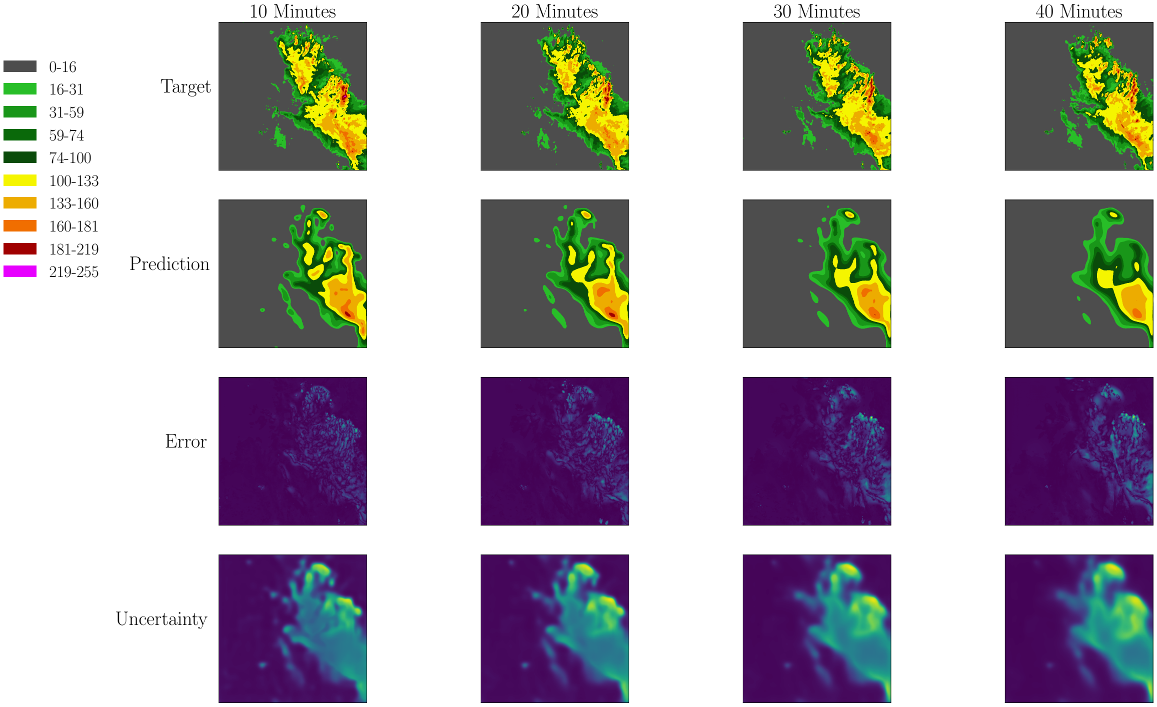

Figure 7: Plot of error and uncertainties for EDL model

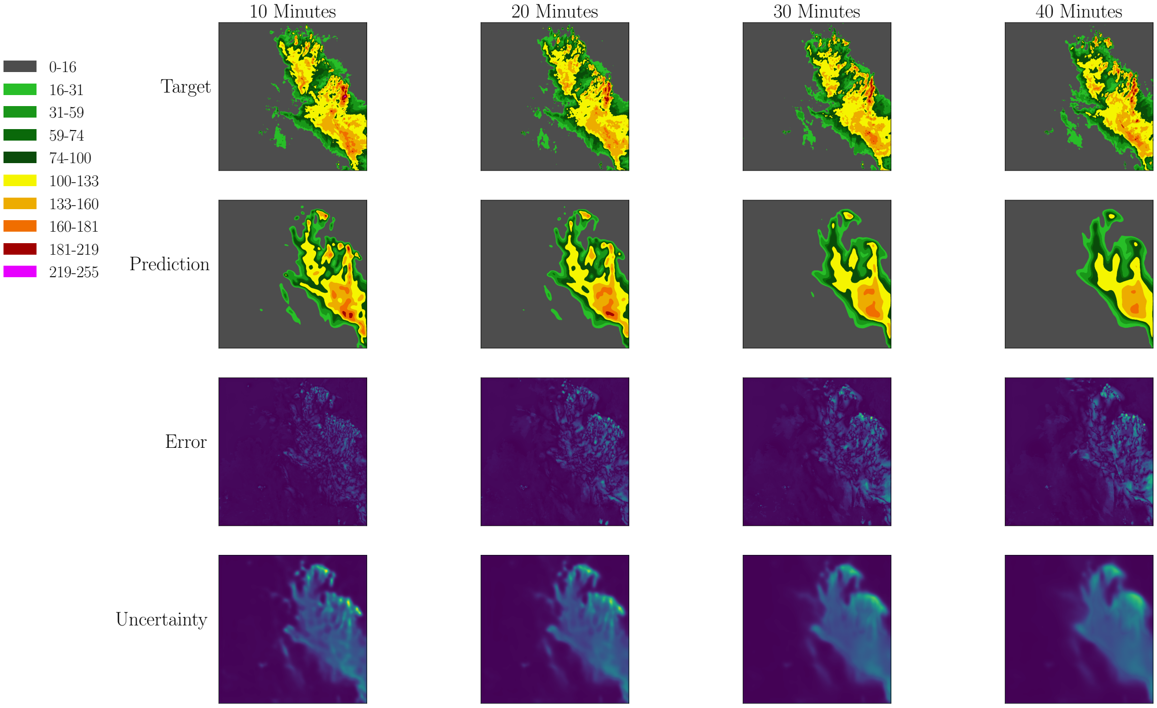

Figure 8: Plot of error and uncertainties for P-EDL model

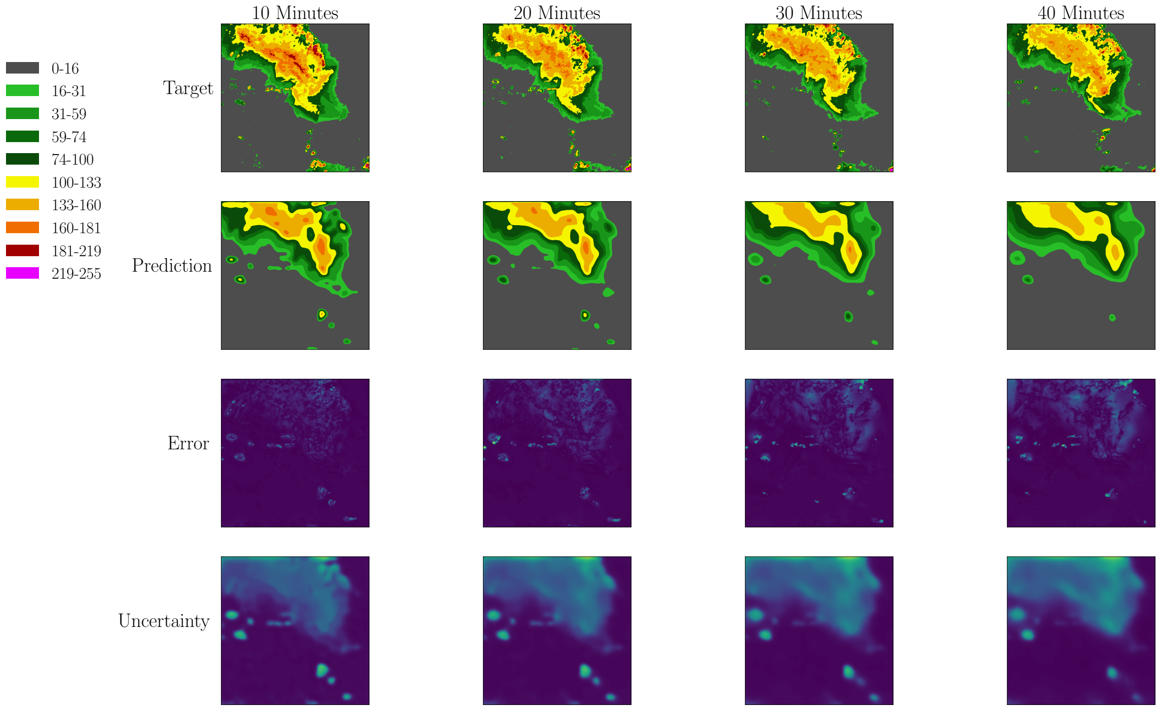

Figure 9: Plot of error and uncertainties for EDL model

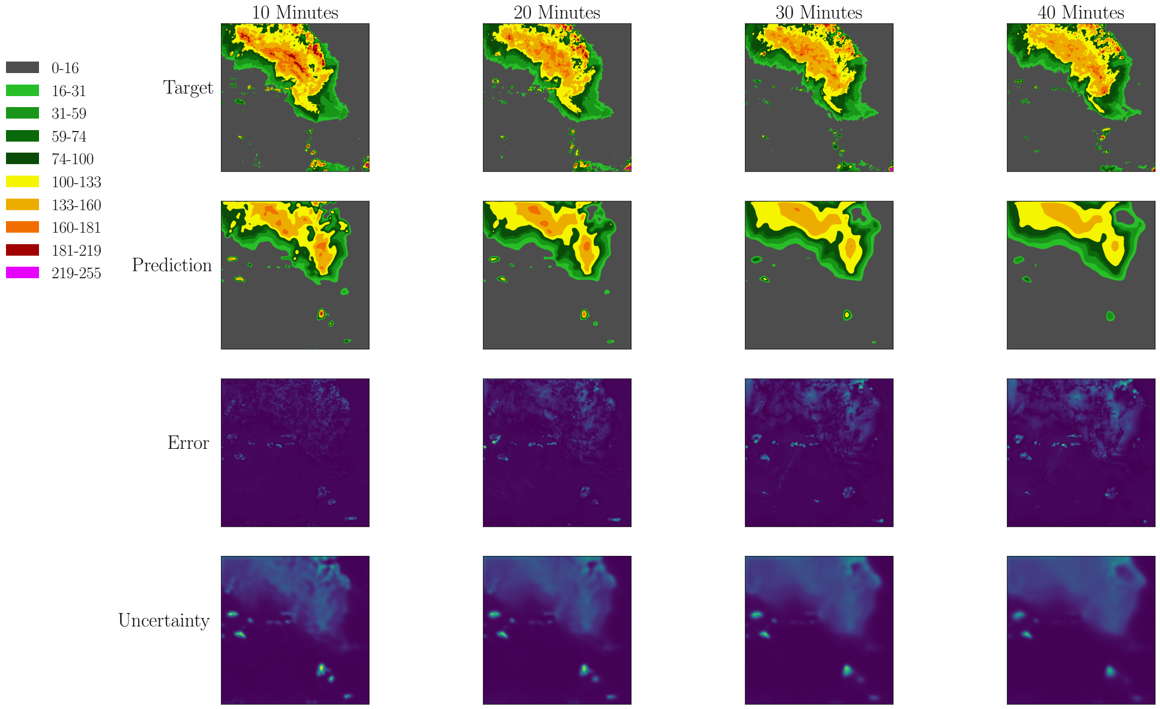

Figure 10: Plot of error and uncertainties for P-EDL model