Uncertainty-aware segmentation for rainfall prediction post processing

Uncertainty-aware segmentation for rainfall prediction post processing

Simone Monaco
simone.monaco@polito.it
Department of Control and Computer Engineering, Politecnico di Torino
Torino, Italy

Luca Monaco
luca.monaco@polito.it
Department of Environmental Engineering, Politecnico di Torino
Torino, Italy

Daniele Apiletti
daniele.apiletti@polito.it
Department of Control and Computer Engineering, Politecnico di Torino
Torino, Italy

ABSTRACT

Accurate precipitation forecasts are crucial for applications such as flood management, agricultural planning, water resource allocation, and weather warnings. Despite advances in numerical weather prediction (NWP) models, they still exhibit significant biases and uncertainties, especially at high spatial and temporal resolutions. To address these limitations, we explore uncertainty-aware deep learning models for post-processing daily cumulative quantitative precipitation forecasts to obtain forecast uncertainties that lead to a better trade-off between accuracy and reliability. Our study compares different state-of-the-art models, and we propose a variant of the well-known SDE-Net, called SDE U-Net, tailored to segmentation problems like ours. We evaluate its performance for both typical and intense precipitation events.

Our results show that all deep learning models significantly outperform the average baseline NWP solution, with our implementation of the SDE U-Net showing the best trade-off between accuracy and reliability. Integrating these models, which account for uncertainty, into operational forecasting systems can improve decision-making and preparedness for weather-related events.

KEYWORDS

Uncertainty-aware deep learning, rainfall prediction,

ACM Reference Format:
Simone Monaco, Luca Monaco, and Daniele Apiletti. 2024. Uncertainty-aware segmentation for rainfall prediction post processing. In *ACM KDD 2024 Workshops, Workshop on Uncertainty Reasoning and Quantification in Decision Making, August 25–29, 2024, Barcellona, Spain.* ACM, New York, NY, USA, 6 pages.

1 INTRODUCTION

Accurate precipitation forecasts are essential for flood management, agricultural planning, water resource allocation, and weather warnings. Despite significant advancements in numerical weather prediction (NWP) models, these models still exhibit biases and uncertainties, especially at high spatial and temporal resolutions. This is due to the complex, nonlinear nature of atmospheric processes and inherent approximations in NWP models [4, 25]. The direct model output (DMO) of NWPs is highly sensitive to initial conditions, boundary conditions, and parameterization schemes (e.g., orography). Consequently, predictions are incomplete without a characterization of the associated uncertainty [7, 8]. For instance, forecast uncertainty is crucial for the Italian Civil Protection in issuing localized severe weather warnings.

Post-processing techniques have been developed to mitigate NWP limitations and improve prediction reliability. Traditional statistical methods like model output statistics (MOS) and ensemble model output statistics (EMOS) have been somewhat successful, but often fail to capture the complexity of precipitation patterns and uncertainties [14, 27].

Recently, machine learning (ML) has shown remarkable results in improving weather forecasts by processing large datasets and recognizing complex patterns that conventional methods struggle with [2, 28].

Our contributions focus on enhancing the reliability and consistency of rainfall forecast uncertainty estimates through post-processing daily cumulative quantitative precipitation forecasts (QPF) from NWPs in northwestern Italy. We aim to improve prediction accuracy while ensuring reliable uncertainty estimates in precipitation forecasts. By reinterpreting rainfall estimation as an image segmentation task, we explore the application of various deep-learning approaches to develop a post-processing tool that integrates forecasts from multiple NWP models. Alongside state-of-the-art solutions, we introduced SDE U-Net, a variant of SDE-Net. [17], specifically designed for segmentation tasks.

This multi-model approach leverages the strengths of individual numerical models, combining them to enhance overall forecast accuracy and reliability [12]. We then comprehensively evaluated the proposed algorithms, particularly focusing on uncertainty estimation for typical and intense weather events. Our analysis addresses the accuracy-reliability tradeoff, balancing confidence in model predictions with the risk of forecasts missing the physical outcomes.

The post-processing systems investigated in this work can be integrated into operational forecasting systems, leading to more informed decision-making and better preparation for weather-related events.

2 BACKGROUND

Uncertainty can have different sources. In a machine learning context, the definitions of aleatoric and epistemic uncertainties help us understand and manage the limitations and reliability of our models' predictions. Aleatoric uncertainty is due to inherent noise in the data: this type of uncertainty is present in the observations and cannot be reduced even if we collect more data. It arises from

ACM KDD '24 Workshops, August 25–29, 2024, Barcellona, Spain

the natural variability in the data generation process. Epistemic uncertainty reflects the model's uncertainty about its predictions due to insufficient training data or limited model capacity [18].

2.1 Related Works

One of the most popular research directions for quantifying uncertainty in neural networks involves Bayesian neural networks (BNNs) [20, 21], which quantify prediction uncertainty by imposing probability distributions over model parameters instead of using point estimates. While BNNs provide a principled method for quantifying uncertainty, the exact derivation of parameter posteriors is often computationally difficult, especially for large input data sets, such as in computer vision tasks.

Among the non-Bayesian approaches, a prominent method in this category is model ensembling [15], in which multiple deep neural networks (DNNs) with different initialization are trained and statistics on their predictions are generated for uncertainty estimation. However, training an ensemble of DNNs can be prohibitively expensive.

Other non-Bayesian methods [10] often mix aleatory uncertainty with epistemic uncertainty. Separating these two sources of uncertainty is crucial for many tasks [1]. SDE-Net [17] addresses this problem by introducing a Brownian motion term into the network architecture to capture epistemic uncertainty and view DNN transformations as state evolution in a stochastic dynamical system. However, this architecture is demonstrated on simple classification and regression tasks with tabular data and cannot be directly applied to segmentation tasks and rainfall prediction without modifications.

Several studies have used Monte Carlo (MC) dropout to estimate uncertainty. Wang et al. [31] analyzed the epistemic and aleatory uncertainty for CNN-based medical image segmentation at both pixel and structural levels.

To our knowledge, only a few works have provided estimates of uncertainties in QPF post-processing. Moosavi et al. [22] have recently applied machine learning strategies to estimate and predict NWP errors in precipitation forecasting. Unfortunately, it is specific to the Weather Research and Forecasting (WRF) model and may not generalize well to other weather models.

3 METHODS

We can formulate our task below with a double interpretation. In the *deterministic interpretation*, given a true precipitation map P for a given event and a set of n *imperfect* predictions $\{P_i\}_{i=1,\dots n}$ which are results of as many different NWP models, our deep learning algorithm — represented as a parametric function of weights θ — must produce an output $\hat{P}$ of the form

$$\hat{P} = f(\{P_i\}; \theta) \tag{1}$$

such that the distance function

$$\mathcal{L} = ||P - \hat{P}||_2 \tag{2}$$

is minimized. This equation presents an L_2 loss function, but other alternatives can be employed as needed.

Alternatively, from a *probabilistic* point of view, we can think of the NWP outcomes P_i as different i.i.d. samples from a distribution of a stochastic process of the form

$$P_i = P + \delta p_i \tag{3}$$

Where the δp_i represents the epistemic error provided by each numerical model. In this framework, we expect an uncertainty in the model prediction $\hat{P}$ due to the type of input it was trained on. At the same time, we expect some aleatoric uncertainty due to inherent measure errors in observational data. In this work, we will not directly distinguish between the two and will just provide overall forecast uncertainty estimates where we consider both sources of error.

Conventional deep learning models are generally used deterministically, providing no access to prediction uncertainty. To address this limitation, we propose to reformulate the problem by replacing the parametric model f with a variant that can produce a distribution of outcomes instead of a single value. In other words, the model prediction can be represented as a sample from this distribution:

$$\hat{P} \sim \tilde{f}(\{P_i\}; \theta) \tag{4}$$

Where $\tilde{f}$ represents the variational model. Given $\bar{Y} = \{\hat{P}_i\}_n$ a set of n samples from the predictive distribution, we can define the prediction intervals (PIs) with a confidence level of $\gamma \in [0, 1)$ as the range $[l(\bar{Y}), u(\bar{Y})]$ such that the probability $\mathcal{P}\left(l(\bar{Y}) < \hat{P}_{n+1} < u(\bar{Y})\right) = \gamma$, which indicate the expected error between the prediction and the actual targets. A large PI indicates greater uncertainty in the model's predictions. While the actual precipitation value is likely to be within the specified interval, the predictions may not be very accurate. Essentially, a large PI indicates that the model has less confidence in its predictions, reflecting greater variability in the input data or inherent challenges in the prediction process.

Conversely, a small PI indicates a higher confidence in the model's predictions, which suggests that the actual precipitation value is likely to be very close to the predicted value. However, this also increases the risk that the actual values will lie outside these PIs. The optimal PI range, therefore, depends heavily on the practical application and is a tradeoff between accuracy and reliability.

In the context of rainfall prediction, NWP simulations P_i typically exhibit a large PI due to varying mathematical assumptions in the different models. While this broad PI is beneficial for capturing intense meteorological events, it can also lead to excessive uncertainty. Ideally, a refined model should reduce this range while maintaining sufficient width to capture significant weather events effectively.

3.1 Case study

Our study aims to estimate forecast uncertainties in daily cumulative QPF over northwestern Italy, specifically focusing on the Piemonte and Valle d'Aosta regions over 24 hours. These areas present a particular challenge for precipitation forecasting due to their varied topography, significantly influencing local precipitation patterns.

To address this task, we compiled a dataset of gridded daily cumulative precipitation observations from ground stations provided by ARPA Piemonte, covering the Area of Interest (AoI) with a spatial resolution of approximately 12 km. These observations, namely the P_is, are interpreted as images of size $L \times W$, with each pixel

being the precipitation within the associated land area. For each real observation, n NWP outcomes are collected and gridded to match the shape of the ground truth, producing an image of size $L \times W \times n$ when stacked together along the channel axis.

3.2 Deep learning architectures

Within this framework, the problem can be phrased as a segmentation task. We chose a U-Net architecture [26] as our deterministic backbone network. Despite the availability of many newer alternatives, U-Net remains extremely popular in fields such as medical imaging, remote sensor analysis, and diffusion models [23, 29]. Its skip-connected encoder-decoder structure is particularly well suited for capturing both local and global contexts, making it ideal for our task, for which we have experimentally found that other more complex architectures do not yield remarkable results. However, it is worth noting that our choice of U-Net is not crucial for the subsequent analysis. All the model changes we will present can also be applied to other segmentation backbones.

Based on this, we have developed several models for segmentation under uncertainty that incorporate the best-known strategies from the literature to achieve this property for different tasks. In the following sections, we briefly introduce these models and highlight our contributions to the development of some of them.

3.2.1 Monte Carlo Dropout U-Net. Henceforth MCD U-Net, this approach enhances the backbone model with a Monte Carlo Dropout (MCD) strategy [24]. Dropout, originally introduced as a regularization procedure, involves randomly discarding a subset of neurons during training to prevent overfitting and improve the generalization of the model. In MCD, this concept is extended to the testing phase to estimate uncertainty, as the dropout at the time of inference can be considered as a Bayesian approximation. Dropout layers are applied during inference, and multiple forward passes are performed to generate a distribution of predictions. The variance of these predictions is then a measure of the uncertainty of the model.

3.2.2 Deep Ensemble U-Net. Henceforth Ens U-Net, in this technique, several U-Net models are trained independently of each other with different initializations [15]. As with the previous method, this approach also leads to variability in the model results, although the number of trained models limits the possible different output patterns.

3.2.3 SDE U-Net. SDE-Net was recently proposed by Kong et al. [17] to integrate Stochastic Differential Equations (SDEs) into deep learning models for capturing uncertainty. Neural networks can be viewed as continuous-time transformations of input dynamics, with model epistemic uncertainty accessed by viewing this process as a stochastic dynamical system governed by the following stochastic differential equation:

$$dx_t = f(x_t, t; \theta_f)dt + g(x_0; \theta_g)dW_t \tag{5}$$

Here, the diffusion term g modulates the Brownian motion dW_t, representing the stochastic component of the process. The parametric functions $f(\cdot; \theta_f)$ and $g(\cdot; \theta_g)$ are two neural networks trained to model aleatoric and epistemic uncertainty, respectively. The training strategy ensures that g provides a small variance for data within the training distribution and a large variance for data outside it. This is obtained by addressing the following objective function:

$$\min_{\theta_f} \mathbb{E}\left[\mathcal{L}(x_T)\right] + \min_{\theta_g} \mathbb{E}_{x_0}\left[g(x_0; \theta_g)\right] + \max_{\theta_g} \mathbb{E}_{\tilde{x}_0}\left[g(\tilde{x}_0; \theta_g)\right], \tag{6}$$

where $\mathcal{L}$ is the task-dependent loss function enforcing stochastic process' terminal outcome x_T to approach the target prediction and $\tilde{x}0$ is an out-of-distribution sample obtained by adding Gaussian noise to the initial state x_0 sampled from the training data.

The original implementation of SDE-Net develops the input-output system over the time interval $[0, T]$ using an Euler-Maruyama scheme, where the two components of the equation 5 are added iteratively with a fixed step size. This allows using the same networks at each time step, reducing the overall number of weights.

Extending this strategy to the U-Net architecture is a challenge because the main advantage of U-Net lies in its networked encoder-decoder structure. In U-Net, the input signal going into each encoder block generates an output that serves as the input for the next encoder block and is also passed to the corresponding decoder block via skip connections. These blocks have different input and output channels, which makes it impossible to share weights between them. To incorporate the SDE-Net strategy, we set the number of time splits to match the number of encoder blocks. For each encoder block, we construct a diffusion block placed at each skip connection and simulate an integration step at each encoder-decoder exchange.

This network is trained using the strategy proposed in [17] to assign higher uncertainty to out-of-distribution inputs, enabling effective uncertainty quantification in segmentation tasks while preserving the essential U-Net structure.

3.3 Experimental design and Validation metrics

To measure the benefits of uncertainty-aware architectures in rainfall prediction tasks, we train the models to reconstruct precipitation maps from different *typical* events. In contrast, we also collect a set of events labelled as *intensive* and separated from the training data. Intensive events are all those where the maximum recorded precipitation within the RoI (Region of Interest) exceeds the 99th percentile for the corresponding season. Further insights are given in the following section. Based on this separation, we expect a deep learning model to perform better when evaluated on typical events, while the performance degrades for intense events. We compare the uncertainty provided by a PoorMan's Ensemble (average of NWP forecasts), which is our benchmark, with the forecast uncertainty from each considered machine learning model. To get a basic uncertainty estimate, we use normalized rMSE, while to quantify the trade-off between accuracy and reliability, we introduce a coverage-length-based criterion (CLC) as defined in [19]

$$CLC = \frac{NMPIL}{\sigma\left(PICP, \eta, \mu\right)}, \tag{7}$$

where σ is a sigmoid function with scaling parameter η and translation parameter μ:

$$\sigma\left(PICP, \eta, \mu\right) = \frac{1}{1 + e^{-\eta(PICP - \mu)}} \tag{8}$$

We aim to achieve low values of Normalized Mean Prediction Interval Length (NMPIL), as it indicates a narrower spread in ensemble predictions, which we seek to minimize for more meaningful and useful predictions. However, reducing NMPIL negatively affects the

coverage of Prediction Intervals (PIs), resulting in an undesirable number of predictions falling outside the PIs. To address this issue, we aim for high PI Coverage Probability (PICP) values, which measure the proportion of target values within the prediction interval. Consequently, we strive for the smallest possible values of CLC. The parameter η controls the penalty when PICP falls below the minimum acceptable level μ.

Ideally, the threshold value of acceptability μ should be as close as possible to 1, so we set it to a reasonably high value, namely $\mu = \gamma = 0.95$. In [5] the parameter η is explored in the context of neural networks training in order to study learning sensitivity and dynamics, leading to a useful range $1 < \eta < 10$: this contribution can be applied also in other contexts such as CLC, so we examine the behaviour of CLC as a function of η ranging from 0 to 12, which slightly extends the suggested range. Of course, predictions falling outside a PI with $\mu = 0.95$ should be strongly penalized, so we are particularly interested in the CLC values for high η (i.e. around 10). Accordingly, we provide tabular values for rMSE, PICP, NMPIL, and CLC with $\eta = 12$, the maximum value considered in our analysis.

MCD U-Net and SDE U-Net use 20 sampled predictions, while Ens U-Net (3.2.2) is based on 5 repetitions, i.e., as many ensemble models. This approach estimates forecast uncertainty for each model, reflecting epistemic error. We then repeat the process in a 9-fold cross-validation to ensure statistical significance, accessing aleatoric uncertainty. This involves training the models on nine different training-validation-test subsets, derived using weather physics considerations detailed in subsection 4.1. Uncertainty estimates for each validation metric VM are provided as:

$$VM = VM_{repetitions} \pm VM_{9-fold\ CV}. \tag{9}$$

4 EXPERIMENTS

4.1 Dataset building

As previously mentioned, the dataset includes observations from ground stations provided by ARPA Piemonte, preprocessed using optimum interpolation [13] to generate images on a fixed grid. These observations span from 1957 to the present, providing a continuous and comprehensive record of precipitation across various meteorological conditions. The dataset also includes precipitation forecasts from four NWP models: BOLAM-CNR [6], ECMWF-IFS [11], COSMO-2I [3], and COSMO-5M [9].

Events were classified as "intense" if their spatial maximum precipitation exceeded the seasonal 99th percentile, with thresholds of 64.58mm in winter, 95.71mm in spring, 93.26mm in summer, and 140.40mm in autumn. This classification resulted in 436 events, 40 of which were marked as intense and set aside during the training phase.

For typical events, we applied K-means clustering based on the variability-average plane to categorize events into convective, stratiform, and intermediate types. These types are characterized by high spatial variability and low spatial average precipitation, low spatial variability and high spatial average precipitation, and intermediate characteristics, respectively [16, 30]. The dataset was then divided into nine distinct training-validation-test subsets, ensuring uniform event type distribution across all subsets. The code for our experiments is available online [1], while the dataset can be shared upon request.

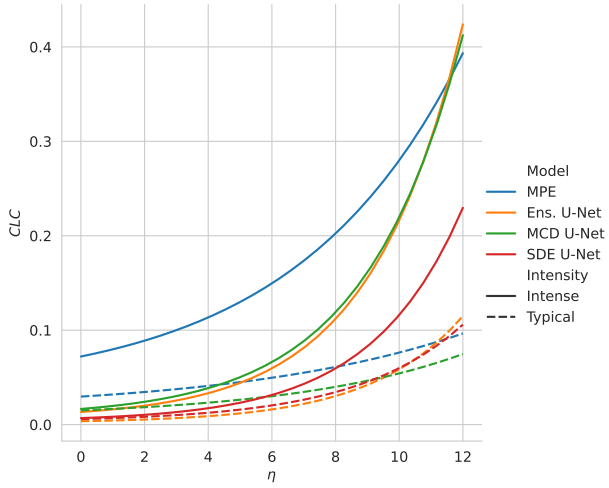

Figure 1: CLC Score of the model over the parameter η

4.2 Results

Figure 1 displays the behaviour of CLC for $\mu = 0.95$ and η values ranging from 0 to 12 for typical and intense events, comparing all deep learning models to the average of the weather models (PoorMan's Ensemble, PME). Error bars are omitted for readability. Smaller CLC values indicate a better trade-off between accuracy and reliability, achieved through low NMPIL and high PCIP values, particularly at higher η values. Table 1 summarizes the results of our analysis in terms of CLC with $\mu = 0.95$ and $\eta = 12$, alongside rMSE, PICP, and NMPIL for all uncertainty-aware models compared to PME.

As expected, rMSE is higher for intense events than for typical events, with all deep learning models significantly outperforming PME. The low uncertainty in rMSE estimates highlights the model with the best rMSE. Ens U-Net achieves the lowest rMSE for typical events ($8.15 \cdot 10^{-3}$), while SDE U-Net achieves the lowest for intense events ($2.637 \cdot 10^{-2}$), indicating higher prediction accuracy.

The PICP column shows that PME has better percentage coverage than the learning models by 10 to 15%, but at the cost of much wider prediction intervals, as indicated by the NMPIL column. PME has NMPIL values of $1.489 \cdot 10^{-2}$ for typical events and $3.605 \cdot 10^{-2}$ for intense events, compared to the deep learning models' range of $2 \cdot 10^{-3}$ to $8 \cdot 10^{-3}$ for both event types. This suggests that PME predictions are reliable but lack accuracy.

However, the trade-off between accuracy and reliability differs for typical and intense events when considering CLC at $\eta = 12$. For typical events, CLC indicates no substantial advantage for deep learning models over PME ($0.09 < CLC < 0.11$ for PME, Ens U-Net, and SDE U-Net), although MCD U-Net shows a slight improvement (0.065). For intense events, SDE U-Net has the lowest CLC value (0.229), with the differences between PME, MCD U-Net, and Ens

[1] https://github.com/simone7monaco/rainfall-prediction

	# Parameters (M)	rMSE (×100) ↓		PICP ↑		NMPIL (×100) ↓		CLC ($\mu = 0.95, \eta = 12$) ↓	
		Typical	Intense	Typical	Intense	Typical	Intense	Typical	Intense
PME	-	1.156	3.152	**0.808**	**0.759**	1.483	3.605	0.096	0.393
Ens. U-Net	5×31.0	**0.779±0.113**	2.746±0.037	0.626±0.032	0.608±0.005	**0.222±0.074**	0.681±0.053	0.110±0.036	0.419±0.020
MCD U-Net	31.0	0.834±0.091	2.849±0.063	0.787±0.027	0.628±0.004	0.791±0.074	0.804±0.048	**0.065±0.017**	0.393±0.026
SDE U-Net	12.4	0.815±0.167	**2.637±0.016**	0.665±0.016	0.602±0.001	0.299±0.011	**0.345±0.019**	0.095±0.014	**0.229±0.009**

Table 1: rMSE, PICP, NMPIL e CLC at $\eta = 12$ in deep learning models vs PoorMan's Ensemble. Up and down arrows indicate whether the best value is the higher or the lower.

U-Net being negligible. While SDE-UNet does not have the highest PICP, its primary advantage is the small prediction intervals, as reflected in NMPIL, especially for intense events ($3.45 \cdot 10^{-3}$). This results in a highly favourable accuracy-reliability trade-off, as shown by CLC.

For $8 < \eta < 12$, PME consistently shows higher CLC values for typical events compared to deep learning models, though the difference is minor. SDE U-Net and Ens U-Net perform comparably or slightly better than MCD U-Net around $\eta = 9$. For intense events, PME consistently underperforms against the deep learning models. MCD U-Net and Ens U-Net have similar CLC values, but SDE U-Net demonstrates the best accuracy-reliability tradeoff.

Overall, these results highlight the effectiveness of deep learning models, particularly the SDE U-Net, in providing accurate and precise rainfall predictions while maintaining a reasonable level of uncertainty quantification.

5 CONCLUSIONS

To summarise, our study demonstrates the effectiveness of deep learning solutions to improve the accuracy and reliability of NWP post-processing systems for precipitation forecasts. By evaluating both typical and intense precipitation events, we found that all deep learning models significantly outperformed the average baseline NWP solution, with our implementation of SDE-UNet showing the best trade-off between accuracy and reliability.

Integrating these models, which account for uncertainty, into operational forecasting systems can improve decision-making and better preparation for weather-related events. Future work will focus on refining these models and exploring alternatives to achieve more comprehensive results in predicting precipitation while accounting for uncertainty.

ACKNOWLEDGMENTS

This work is part of the project NODES, funded by the Italian MUR (Ministry of University and Research) under M4C2 1.5 of the PNRR (National Plan for Recovery and Resilience) with grant agreement no. ECS00000036. The SmartData@PoliTO research centre of Politecnico di Torino, Italy has partially funded this work.

REFERENCES

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

[2] ACM 2018. *DeepSD: Generating high resolution climate change projections through single image super-resolution.* ACM.

[3] Michael Baldauf, Axel Seifert, Jochen Förstner, Detlev Majewski, Matthias Raschendorfer, and Thorsten Reinhardt. 2011. Operational convective-scale numerical weather prediction with the COSMO model: Description and sensitivities. *Monthly Weather Review* 139, 12 (2011), 3887–3905.

[4] Peter Bauer, Alan Thorpe, and Gilbert Brunet. 2015. The quiet revolution of numerical weather prediction. *Nature* 525, 7567 (2015), 47–55.

[5] Christopher M. Bishop. 1995. *Neural Networks for Pattern Recognition.* Oxford University Press.

[6] A Buzzi, M Fantini, P Malguzzi, and F Nerozzi. 1994. Validation of a limited area model in cases of Mediterranean cyclogenesis: surface fields and precipitation scores. *Meteorology and Atmospheric Physics* 53, 3 (1994), 137–153.

[7] Julie Demargne, Limin Wu, Satish K Regonda, James D Brown, Haksu Lee, Minxue He, Dong-Jun Seo, Robert Hartman, Henry D Herr, Mark Fresch, et al. 2014. The science of NOAA's operational hydrologic ensemble forecast service. *Bulletin of the American Meteorological Society* 95, 1 (2014), 79–98.

[8] Julie Demargne, Limin Wu, Satish K. Regonda, James D. Brown, Haksu Lee, Minxue He, Dong-Jun Seo, Robert Hartman, Henry D. Herr, Mark Fresch, John Schaake, and Yuejian Zhu. 2014. The Science of NOAA's Operational Hydrologic Ensemble Forecast Service. *Bulletin of the American Meteorological Society* 95, 1 (2014), 79 – 98. https://doi.org/10.1175/BAMS-D-12-00081.1

[9] Georg Doms and Michael Baldauf. 2018. *A Description of the Nonhydrostatic Regional COSMO Model. Part I: Dynamics and Numerics.* Technical Report. COSMO Technical Report, Deutscher Wetterdienst.

[10] J. et al. Dy (Ed.). 2018. *Bias-Reduced Uncertainty Estimation for Deep Neural Classifiers.*

[11] ECMWF. 2016. *IFS Documentation CY43R1.* Technical Report. ECMWF Technical Documentation.

[12] David J Gagne, Amy McGovern, and Ming Xue. 2014. Machine learning enhancement of storm-scale ensemble precipitation forecasts. *Weather and Forecasting* 29, 4 (2014), 1024–1043.

[13] Lev S Gandin. 1963. *Objective Analysis of Meteorological Fields.* Gidrometeorizdat, Leningrad.

[14] Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E Raftery. 2007. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 69, 2 (2007), 243–268.

[15] I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). 2017. *Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles.* Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf

[16] Robert A Houze. 1997. Stratiform precipitation in regions of convection: A meteorological paradox? *Bulletin of the American Meteorological Society* 78, 10 (1997), 2179–2196.

[17] International Machine Learning Society (IMLS) 2020. *SDE-Net: Equipping Deep Neural Networks with Uncertainty Estimates.* International Machine Learning Society (IMLS).

[18] Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in Bayesian deep learning for computer vision?. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17).* Curran Associates Inc., Red Hook, NY, USA, 5580–5590.

[19] Abbas Khosravi, Saeid Nahavandi, and Doug Creighton. 2010. Construction of optimal prediction intervals for load forecasting problems. *IEEE Transactions on Power Systems* 25, 3 (2010), 1496–1503.

[20] R. P. Lippmann, J. Moody, and D. Touretzky (Eds.). 1990. *Transforming Neural-Net Output Levels to Probability Distributions.* Vol. 3. Morgan-Kaufmann. https://proceedings.neurips.cc/paper_files/paper/1990/file/7eacb532570ff6858afd2723755ff790-Paper.pdf

[21] David JC MacKay. 1992. A practical Bayesian framework for backpropagation networks. *Neural computation* 4, 3 (1992), 448–472.

[22] Azam Moosavi, Vishwas Rao, and Adrian Sandu. 2021. Machine learning based algorithms for uncertainty quantification in numerical weather prediction models. *Journal of Computational Science* 50 (2021), 101295.

[23] W. Peebles and S. Xie. 2023. Scalable Diffusion Models with Transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV).* IEEE Computer

Society, Los Alamitos, CA, USA, 4172–4182. https://doi.org/10.1109/ICCV51070.2023.00387

[24] PMLR 2016. *Dropout as a Bayesian approximation: Representing model uncertainty in deep learning*. PMLR.

[25] Stephan Rasp, Peter D Dueben, Sebastian Scher, Jonathan A Weyn, Soukayna Mouatadid, and Nils Thuerey. 2020. WeatherBench: A benchmark dataset for data-driven weather forecasting. *Geoscientific Model Development* 13, 3 (2020), 1199–1210.

[26] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi (Eds.). Springer International Publishing, Cham, 234–241.

[27] Michael Scheuerer and Thomas M Hamill. 2015. Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Monthly Weather Review* 143, 4 (2015), 1321–1334.

[28] Martin G Schultz, Hao He, and Florian Kleinert. 2021. Can deep learning beat numerical weather prediction? *Philosophical Transactions of the Royal Society A* 379, 2194 (2021), 20200097.

[29] Nahian Siddique, Sidike Paheding, Colin P. Elkin, and Vijay Devabhaktuni. 2021. U-Net and Its Variants for Medical Image Segmentation: A Review of Theory and Applications. *IEEE Access* 9 (2021), 82031–82057. https://doi.org/10.1109/ACCESS.2021.3086020

[30] Soroosh Sorooshian, Kuolin Hsu, Xiaogang Gao, Hoshin V Gupta, Bahram Imam, and Don Braithwaite. 2002. Evaluation of PERSIANN system satellite–based estimates of tropical rainfall. *Bulletin of the American Meteorological Society* 83, 1 (2002), 63–70.

[31] Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. 2019. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing* 338 (2019), 34–45.