Environmental Data Science (2024), 3: e45, 1–23
doi:10.1017/eds.2024.37

CAMBRIDGE UNIVERSITY PRESS

APPLICATION PAPER

Cloudy with a chance of uncertainty: autoconversion rates forecasting via evidential regression from satellite data

Maria Carolina Novitasari[1], Johannes Quaas[2,3] and Miguel R. D. Rodrigues[1]

[1]Department of Electronic and Electrical Engineering, University College London, London, United Kingdom
[2]Leipzig Institute for Meteorology, Universität Leipzig, Leipzig, Germany
[3]ScaDS.AI - Center for Scalable Data Analytics and AI, Leipzig, Germany
Corresponding author: Maria Carolina Novitasari; Email: maria.novitasari.20@ucl.ac.uk

Received: 30 January 2024; **Revised:** 31 August 2024; **Accepted:** 27 September 2024

Keywords: Autoconversion rates; evidential regression; uncertainty quantification; aerosol-cloud interaction; precipitation formation

Abstract

High-resolution simulations such as the ICOsahedral Non-hydrostatic Large-Eddy Model (ICON-LEM) can be used to understand the interactions among aerosols, clouds, and precipitation processes that currently represent the largest source of uncertainty involved in determining the radiative forcing of climate change. Nevertheless, due to the exceptionally high computing cost required, this simulation-based approach can only be employed for a short period within a limited area. Despite the potential of machine learning to alleviate this issue, the associated model and data uncertainties may impact its reliability. To address this, we developed a neural network (NN) model powered by evidential learning, which is easy to implement, to assess both data (aleatoric) and model (epistemic) uncertainties applied to satellite observation data. By differentiating whether uncertainties stem from data or the model, we can adapt our strategies accordingly. Our study focuses on estimating the autoconversion rates, a process in which small droplets (cloud droplets) collide and coalesce to become larger droplets (raindrops). This process is one of the key contributors to the precipitation formation of liquid clouds, crucial for a better understanding of cloud responses to anthropogenic aerosols and, subsequently, climate change. We demonstrate that incorporating evidential regression enhances the model's credibility by accounting for uncertainties without compromising performance or requiring additional training or inference. Additionally, the uncertainty estimation shows good calibration and provides valuable insights for future enhancements, potentially encouraging more open discussions and exploration, especially in the field of atmospheric science.

Impact Statement

This research employs a cutting-edge approach—deep evidential regression—to predict autoconversion rates from satellite data. The distinctive advantage of this methodology lies in its inherent capacity to concurrently assess both data and model uncertainties, all without requiring additional training or inference steps. This not only reduces the overall cost but also significantly increases the credibility of the machine learning model. In addition, evidential regression is easy to implement. By seamlessly incorporating uncertainty estimation into the prediction of autoconversion rates, evidential regression offers a new perspective that goes beyond prediction. This enhanced credibility and improved understanding of uncertainties have the potential to foster greater transparency and trustworthiness in machine learning results, paving the way for a broader discussion, particularly within the domain of atmospheric science.

This research article was awarded Open Data badge for transparent practices. See the Data Availability Statement for details.

1. Introduction

Future climate projections demand a deeper understanding of interactions between aerosols and clouds since they constitute the biggest uncertainty in estimating the radiative forcing of climate change (IPCC, 2021). One way to reduce such uncertainty is to understand the interaction between aerosols, clouds, and precipitation processes.

Aerosols have a dual effect on climate – directly, through interactions with solar radiation involving scattering and absorbing, and indirectly, by modifying cloud properties, thereby impacting Earth's energy budget. Specifically, high aerosol concentrations can reduce the radius of cloud droplets, thereby increasing the cloud albedo (Twomey, 1974). At the same time, smaller cloud droplets reduce precipitation efficiency and prolong cloud lifetime (Albrecht, 1989; Gryspeerdt et al., 2019; Bellouin et al., 2020). This interplay leads to complex feedback mechanisms, making it challenging to accurately quantify the radiative forcing caused by aerosol-cloud interactions. This complexity contributes to uncertainties in our understanding of future climate projections.

One of the high-resolution simulations suitable for investigating the intricate interactions among aerosols, clouds, and precipitation is the high-resolution ICON-LEM (Zängl et al., 2015; Dipankar et al., 2015; Heinze et al., 2017). Nevertheless, due to the exceptionally high computing cost required, it can only be employed for a limited period and geographical area. Specifically, running this simulation model demands around 13 hours on 300 computer nodes, only to generate a single hour of climate data over Germany, incurring a cost of around EUR 100,000 per simulated day (Costa-Surós et al., 2020).

While machine learning has the potential to alleviate the challenges mentioned (Novitasari et al., 2024, attached), it is important to note that predictions made by machine learning models may be affected by noise and model inference errors (Malinin, 2019). In our previous research (Novitasari et al., 2024, attached), we demonstrated that machine learning can effectively predict one of the key processes of precipitation, namely autoconversion rates, directly from satellite data. This approach leverages satellite data, which covers extensive geographical areas and spans over two decades. Because satellite data is both comprehensive and easily accessible, it enables significant cost savings by reducing the need for extensive and costly atmospheric simulation data. However, the degree of uncertainty in a model prediction depends on both the model itself and the quality and availability of data. These uncertainties might be more pronounced for climate forecasts, raising concerns about the model's reliability.

In machine learning, uncertainty can be disentangled into aleatoric and epistemic uncertainties (Der Kiureghian and Ditlevsen, 2009; Hüllermeier and Waegeman, 2021). Aleatoric uncertainty is related to data and can be learned directly from the data. It typically does not depend on the sample size. Meanwhile, epistemic uncertainty is related to the machine learning model or prediction and can be reduced by adding more sample data to the trained model or refining the machine learning model itself.

In agreement with the concepts discussed by Haynes et al. (2023), in machine learning literature, there is occasional interchangeability or confusion between the terms 'aleatoric' and 'stochastic' or 'irreducible,' leading to the perception that aleatoric uncertainty is synonymous with irreducible uncertainty. However, the distinction between aleatoric and epistemic uncertainty in machine learning is context-dependent and hinges on whether the uncertainty is inherent to the data itself or arises from the model's limitations, rather than solely based on the stochastic nature of the uncertainty source.

For instance, enhancing a machine learning (ML) model can reduce epistemic uncertainty without affecting aleatoric uncertainty. Yet, augmenting the dataset with additional features, without increasing its size, might paradoxically decrease aleatoric uncertainty by providing more insights into the inherent noise or randomness in the system, while potentially increasing epistemic uncertainty due to the model's inability to capture more complex relationships without more data. This scenario challenges the notion of aleatoric uncertainty being 'irreducible' in the context of ML, as introducing more informative features can help better characterize and potentially reduce the inherent noise or randomness in the data.

As an example, consider a dataset with sensor measurements. Adding additional features such as environmental conditions or sensor metadata can help better understand and account for the inherent noise or randomness in the sensor measurements, thereby reducing the aleatoric uncertainty associated with the

https://doi.org/10.1017/eds.2024.37 Published online by Cambridge University Press

data. However, this assumes that the additional features are indeed informative and relevant to capturing the inherent noise or randomness in the data. At the same time, these additional features may also increase the complexity of the data-generating process, leading to higher epistemic uncertainty for the model until more data is provided to learn the intricate relationships.

Several approaches for handling uncertainty in ML exist. Bayesian Neural Networks (BNNs) (Kendall and Gal, 2017) are among the most common methods for uncertainty quantification. BNNs incorporate probabilistic priors over the weights and employ sampling techniques to estimate the variance of the output. However, BNN models involve high computational costs due to their iterative sampling during inference. Another prevalent method for uncertainty estimation is deep ensembles (Lakshminarayanan et al., 2017), which offer a simpler implementation compared to BNNs. Nonetheless, they are also computationally very expensive as they require additional training. In practical applications, implementing BNNs or deep ensemble models is often more challenging and involves slower computational training/inference when contrasted with non-Bayesian Neural Networks and models without ensemble methods.

In our study, we advocate for the adoption of deep evidential regression (Amini et al., 2020; Meinert et al., 2023) as an alternative. Deep evidential regression provides a compelling framework for uncertainty modeling without the inherent computational burden associated with Bayesian NNs and ensemble models. Unlike these approaches, evidential deep learning seamlessly integrates uncertainty quantification into the learning process, eliminating the need for extra training or inference steps. In addition, it is easy to implement in practice.

Thus, the aforementioned issues are addressed in this study by employing deep evidential regression and the massive collection of satellite data, providing long-term global spatial coverage up to several decades. Specifically, we aim to address the well-known issue of overconfident predictions in ML algorithms, which frequently overlook the inherent variability and randomness in atmospheric physical processes. For example, traditional models might predict a specific amount of rainfall in a given area, but they often overlook the inherent randomness introduced by sensor noise or measurement errors. Even high-quality sensors can have variability in their readings due to factors such as slight calibration differences or environmental interference. This randomness, known as aleatoric uncertainty, results in predictions that seem more precise than they actually are, as the model might not fully account for the randomness in the sensor data. Additionally, many ML models suffer from epistemic uncertainty, which arises from limitations in the model itself, such as insufficient data to fully capture complex atmospheric dynamics. When models fail to account for these uncertainties, they produce overconfident predictions that can mislead decision-makers, particularly in the context of climate modeling where accurate uncertainty quantification is crucial.

We focus in particular on autoconversion rate estimation since this is one of the key processes in the precipitation formation of liquid clouds, hence crucial to better understanding cloud responses to anthropogenic aerosols and, ultimately, climate change.

Our current work represents a continuation and enhancement of our previous research efforts. Our previous work (Novitasari et al., 2021; Novitasari et al., 2024, attached) also considered the prediction of autoconversion rates from satellite data; however, we have not quantitatively assessed the inherent uncertainty in our previous findings. In our previous research, we proposed predicting autoconversion rates directly from satellite data via a two-stage process: first, we trained, validated, and tested the autoconversion rates using simulation model output (ICON); then, we used the best model to predict the autoconversion rates directly from satellite data (MODIS). We explored various ML models, including random forest (RF), shallow neural network (NN), and deep neural network (DNN), but found that the performance differences among them were not significant. This led us to hypothesize that altering the architecture of the ML model with the same dataset might not substantially improve performance. We suggested that incorporating uncertainty estimation could be a valuable direction for future research to test this hypothesis. In this manuscript, we address this hypothesis by integrating uncertainty quantification methods, thereby enhancing the robustness and interpretability of the predictions. This work offers a computationally effective solution while still obtaining accurate predictions of autoconversion rates with

https://doi.org/10.1017/eds.2024.37 Published online by Cambridge University Press

deep evidential learning, which also quantifies data and model uncertainties without the need for additional training or inference.

Our paper is structured in the following manner: Section 2 provides a detailed exploration of the methodology utilized in our study, covering aspects such as datasets, ML models, procedures employed, and the evaluation of these models. Proceeding to Section 3, we showcase a range of experimental results. This includes hyperparameter selection for the evidential regression models and the uncertainty evaluation, followed by autoconversion rate prediction results using the chosen model on both simulation and satellite data, and then uncertainty estimation for both simulation and satellite data. Finally, in Section 4, we conclude our paper by providing a number of summary remarks and proposing potential areas for future research.

2. Methodology

We present a novel approach for direct autoconversion rate extraction from satellite observation, which includes the estimation of data and model uncertainties via evidential regression. The general framework, as illustrated in Figure 1, involves climate science-based steps for generating training and testing datasets, along with an ML framework incorporating uncertainty quantification methods, as elaborated in the following section. We would like to emphasize again that this work builds upon our previous research (Novitasari et al., 2024, attached). The methodology is very similar to our earlier work, with the key addition of incorporating uncertainty quantification methods.

In general, the left-hand side of the figure illustrates how we extract the input–output pairs used to train and test our ML models from the outputs of the atmospheric simulation models, ICON-LEM and ICON-NWP. Details regarding the datasets are described further in Subsection 2.1.

During the ML framework preparation, we train and validate the models using the evidential regression approach. This involves selecting the best hyperparameters and evaluating the calibration of the uncertainty estimation, as detailed in Subsection 3.1, as well as experimenting with different evidential models, as described in Subsection 3.2.

Once the model training and validation are complete, and we ensure that our uncertainty evaluation shows good calibration, we proceed with model testing and further uncertainty quantification. In the model testing phase, similar to our previous research (Novitasari et al., 2024, attached), we assess the

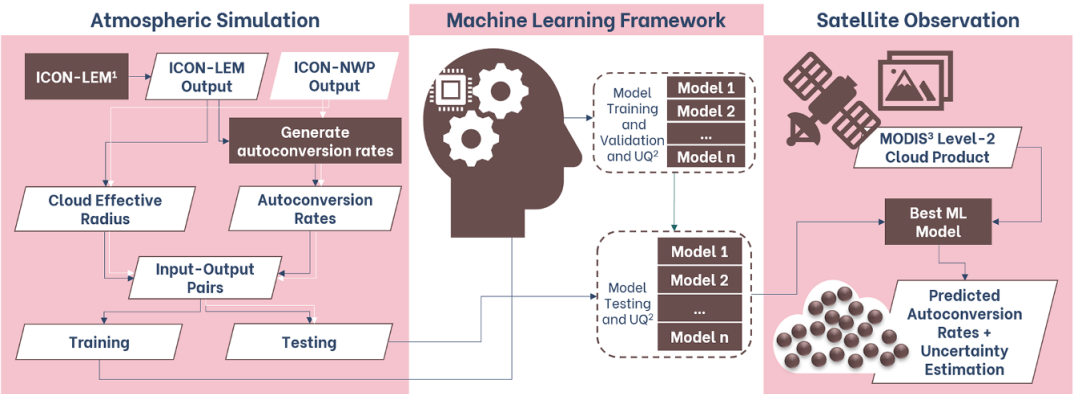

Figure 1. *General framework. The left side of the image illustrates the climate science-based procedures we apply to our dataset to generate input–output pairs for training and testing. The center of the image represents our ML framework, which also includes uncertainty quantification. The right side depicts the satellite observation data we used and the procedure to predict the autoconversion rates from the satellite data, including its inherent uncertainty.* [1] *ICOsahedral Non-hydrostatic Large-Eddy Model;* [2] *Uncertainty Quantification;* [3] *Moderate Resolution Imaging Spectroradiometer.*

https://doi.org/10.1017/eds.2024.37 Published online by Cambridge University Press

performance of our model for predicting autoconversion rates against simulation model output in the initial stage, as detailed in Subsection 3.3. In the second stage, shown on the right-hand side of the figure, we predict autoconversion rates directly from satellite data using our chosen (best) model, as detailed in Subsection 3.4. Although this may appear similar to our previous work, the goal is to demonstrate that incorporating evidential regression enables effective uncertainty quantification without compromising model accuracy. Additionally, we provide uncertainty quantification for both data and model across both stages: simulation and satellite data, as described in Subsection 3.5.

2.1. Data

For our research, we utilize datasets derived from ICON-LEM output over Germany on 02 May 2013, a day characterized by distinct cloud regimes that enable the exploration of various aspects of cloud formation and evolution (Heinze et al., 2017). As described in the study conducted by Costa-Surós et al. (2020), the cloud regimes over ICON-LEM Germany on 2 May 2013 included low-level clouds (Cumulus, Stratocumulus, Stratus), mid-level clouds (Altocumulus, Altostratus, Nimbostratus), high-level clouds (Cirrus, Cirrostratus), and deep convective clouds.

The specific time window of investigation spans from 09:55 UTC to 13:20 UTC with random timesteps. The total number of ICON-LEM images used for training, validation, and testing in this study is five images. We first split the dataset into 80% for training (including validation) and 20% for testing. For this initial split, we divided the data at the image level, where the first four images (based on the timestamp) were used for training and validation, and the last image was used for testing. This method was chosen to simplify the visualization of the test results, and the selection of the last image for testing was arbitrary, as the order of the images is not critical.

After preprocessing, we further split the training set into 80% for training and 20% for validation, with the split performed randomly on the clean data points (pixels). Consequently, the training/validation and testing sets are independent, with no overlap between them. We focus on ICON-LEM with a native resolution of 156 m on the ICON grid, subsequently regridded to a regular 1 km resolution to align with the MODIS resolution. An icosahedral grid divides the surface of a sphere into equilateral triangles, providing a more uniform distribution of grid points compared to latitude-longitude grids, which can become distorted near the poles. The number of clean data points after preprocessing, which covers only cloudy data points, is around 4 million for training and validation.

To further test the performance of our ML model, we use the dataset of ICON numerical weather prediction (ICON-NWP) Holuhraun which was performed over the North Atlantic Ocean around the Holuhraun volcano in Iceland. This dataset spans for a week, from 1 to 7 September 2014, with 1-hour timesteps. During this time, lava was growing quickly, releasing a lot of SO2 into the atmosphere, and clear volcanic plumes were observed, providing a comprehensive examination to further evaluate the effectiveness of our ML models (Kolzenburg et al., 2017; Haghighatnasab et al., 2022). The ICON-NWP simulations for the Holuhraun volcano feature a vertical grid with 75 altitude levels, extending up to 30 km, and a horizontal resolution of about 2.5 km.

The ICON-NWP Holuhraun dataset features diverse meteorological conditions compared to the ICON-LEM Germany. The ICON-NWP comprises a large variety of clouds, including those associated with cold and warm fronts, as well as marine boundary layer clouds. ICON-NWP primarily occurs over the ocean, resulting in a cloud regime that is predominantly marine. In contrast, ICON-LEM mainly occurs over land, leading to a cloud regime that is primarily continental. Both the ICON-LEM and ICON-NWP simulations utilize the two-moment microphysical parameterization proposed by Seifert and Beheng (2006). Both datasets include various features such as temperature, pressure, humidity, cloud effective radius, and cloud optical thickness, among others. However, the final input feature we use is cloud effective radius only, in alignment with our previous study (Novitasari et al., 2024, attached).

Autoconversion rates in both our training and testing datasets are computed using the two-moment microphysical parameterization proposed by Seifert and Beheng (2006). Notably, the autoconversion rates pertaining to cloud tops simulating satellite data are determined based on instances where the cloud optical

https://doi.org/10.1017/eds.2024.37 Published online by Cambridge University Press

thickness (COT), calculated from top to bottom, exceeds 1. The optical thickness serves as a crucial measure indicating the extent to which optical satellite sensors can capture cloud microphysical details.

Furthermore, in relation to the satellite observation data illustrated on the right-hand side of Figure 1, we use the Collection 6 of cloud product level-2 obtained from Moderate Resolution Imaging Spectroradiometer (MODIS) (Platnick et al., 2017, 2018). MODIS is a widely utilized remote sensing instrument in Earth science research, operates on both the Terra and Aqua satellites. Terra's orbit passes the equator from north to south during the morning, while Aqua's orbit traverses from south to north in the afternoon. Terra's satellite typically passes over around 10:30 am local time, and Aqua's overpass occurs at approximately 1:30 pm local time.

2.2. ML models

In this study, our ML models were trained with input derived from ICON-LEM output. To clarify, we use datasets derived from ICON-LEM Germany to train, validate, and test our ML models, and we use datasets derived from ICON-NWP Holuhraun for additional testing only. The details of the training and validation data are outlined in Section 2.1, while the testing scenarios/datasets are explained in Section 2.5. The ML model specifically focuses on the cloud effective radius, a critical parameter in cloud microphysical state, commonly obtained from satellite retrievals as highlighted by previous studies (Platnick et al., 2017; Grosvenor et al., 2018), as input. The resulting output of the ML model, serving as ground truth, comprises autoconversion rates derived from ICON-LEM output.

We split the data into 80% for training and validation, and 20% for testing, as explained in detail in the Data section (Section 2.1). To enhance the model performance, logarithmic transformations are applied to both the input and output variables for normalization. This normalization process effectively addresses data presented with extremely small numerical values, thereby improving the model's stability during training and making the model's predictions more interpretable. Furthermore, the input variables undergo additional normalization using standard scaling techniques, as described below.

Consider a collection of labeled examples consisting of the sets $\mathcal{X}$ and $\mathcal{Y}$, defined as follows:

$$\mathcal{X} = \{x_1, x_2, \ldots, x_N\},$$

where each x_i represents the cloud effective radius (CER) and N represents the size of the collection. Each instance i corresponds to a real-valued target y_i:

$$\mathcal{Y} = \{y_1, y_2, \ldots, y_N\},$$

where y_i represents the autoconversion rate. Here, each x_i and y_i are scalar values (i.e., one-dimensional).

We apply the normalization transformation independently to the single feature x_i in the input data $\mathcal{X}$.

1. Calculate the mean μ and standard deviation σ of the logarithmically transformed feature values across all samples:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} \log_{10}(x_i) \tag{2.1}$$

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (\log_{10}(x_i) - \mu)^2} \tag{2.2}$$

2. Normalize each feature value x_i using the formula:

$$x_i' = \frac{\log_{10}(x_i) - \mu}{\sigma} \tag{2.3}$$

https://doi.org/10.1017/eds.2024.37 Published online by Cambridge University Press

The assessment of each model's performance involves a comprehensive evaluation through the calculation of various metrics. These metrics encompass R^2 (Coefficient of Determination), MAPE (Mean Absolute Percentage Error), RMSPE (Root Mean Squared Percentage Error), PSNR (Peak Signal-to-Noise Ratio), and SSIM (Structural Similarity Index), all of which are applied to the testing data.

R^2 is utilized to assess the overall goodness-of-fit of the model and determine the proportion of variance in the dependent variable that it captures. MAPE measures the average percentage difference between predicted and actual values, providing insights into the accuracy of the model's predictions in a more interpretable form, particularly through percentage-based errors. RMSPE is a variation of MAPE, incorporating a square root to give more weight to larger errors and, consequently, to penalize them more significantly. PSNR measures the ratio between the maximum possible power of a signal and the power of corrupting noise, serving as a metric to evaluate the quality of model predictions. SSIM assesses the similarity between two images, considering luminance, contrast, and structure. It is particularly relevant in our study, which involves image comparison between the ground truth and the model predictions.

Prior to calculating each metric, both the actual output and the predicted output are normalized by transforming them using base 10 logarithms and then scaling them to a range between 0 and 1, ensuring a fair and meaningful evaluation of model performance across diverse datasets and models.

We then trained and tested deep evidential regression models (Amini et al., 2020; Meinert et al., 2023). The NN models are trained to infer the hyperparameters of the evidential distribution by applying evidential priors over the original Gaussian likelihood function. In this study, we exclusively consider NN models since the utilization of NNs, including DNNs, can be paired with evidential learning, facilitating the estimation of both aleatoric and epistemic uncertainties without requiring additional training.

To facilitate comparison, we train our deep evidential regression using two distinct NN architectures. The first model, referred to as the shallow NN or NN, is a basic neural network with a single hidden layer comprising 64 neurons. In contrast, the second model, known as the pyramid-shaped DNN, features a more intricate architecture and a greater number of trainable weights. Both models were also employed in our previous studies (Novitasari et al., 2024, attached; Novitasari et al., 2021).

The DNN architecture of the evidential regression model consists of five fully connected hidden layers, with 1024 nodes in the first layer, 512 nodes in the second layer, 256 nodes in the third layer, 128 nodes in the fourth layer, and 64 nodes in the fifth layer.

For both NN models (NN and DNN), the activation function used at each hidden layer is Leaky ReLU. For the loss function, we use evidential deep regression loss (refer to Section 2.3). Furthermore, the training was performed using Adam's optimizer. The batch size and learning rate are set based on the Keras tuner algorithm (O'Malley et al., 2019). Given our use of evidential deep learning, we modify the final layer of both models to incorporate the evidential component. For enhanced clarity, Figure 2 presents the visualization of the architecture for our NN evidential model. It involves the input of cloud effective radius (r_e) alongside a single hidden layer comprising 64 nodes and 4 outputs, representing γ or

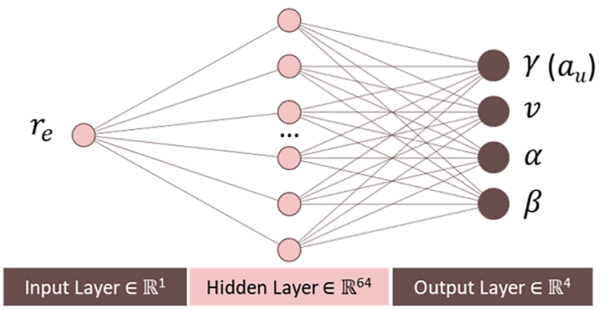

Figure 2. *The architecture of our evidential NN model, with cloud effective radius (r_e) as the input and the autoconversion rate (a_u or γ) as the output, along with three other evidential parameters: υ, α, and β.*

https://doi.org/10.1017/eds.2024.37 Published online by Cambridge University Press

the autoconversion rate (a_u), as well as the remaining evidential parameters utilized for calculating data and model uncertainties (υ, α, β).

We opted for a shallow NN model as our final model due to its ability to produce similar outcomes while utilizing significantly fewer trainable weights compared to a more complex DNN model, as elaborated in Section 3.2. Despite the slightly better outcomes produced by the DNN model, the difference is not statistically significant in our scenario.

In the early stages of our series of experiments, we first selected the evidential hyperparameters and evaluated the uncertainty estimation, as explained in Section 3.1, rigorously assessing their calibration through the implementation of the spread-skill plot (Wilks, 2011; Haynes et al., 2023). To further verify the reliability and calibration of our uncertainty estimates, we also employed the discard test method (Barnes and Barnes, 2021; Haynes et al., 2023). Both uncertainty evaluation methods are discussed in detail in Section 2.4.

For the autoconversion rate estimation stages, we first compared our models' performance against simulation data (as detailed in Section 3.3). We then leveraged our final model (a shallow NN) to directly predict the autoconversion rates from satellite data (see Section 3.4). Finally, extending the scope of our analysis, we calculated the data (aleatoric) and model (epistemic) uncertainties, as explained further in the Experimental Results section (under Section 3.5). This involved not only making predictions but also estimating the associated uncertainties.

2.3. Evidential regression for uncertainty quantification

To capture the uncertainty associated with the predicted autoconversion rates, we adopt the evidential regression framework introduced by Amini et al. (2020). However, we incorporate the modified version as proposed by Meinert et al. (2023).

Amini et al. (2020) approach the regression problem by assuming that the underlying data follows a normal distribution with an unknown mean (μ) and variance (σ^2). This assumption entails assigning a normal prior distribution ($\mathcal{N}$) for the mean, and an inverse-Gamma prior distribution (Γ^{-1}) for the variance, which can be expressed as:

$$\begin{aligned}(y_1,\ldots,y_N) &\sim \mathcal{N}\left(\mu,\sigma^2\right)\\ \mu &\sim \mathcal{N}\left(\gamma,\sigma^2 v^{-1}\right)\\ \sigma^2 &\sim \Gamma^{-1}(\alpha,\beta)\end{aligned} \tag{2.4}$$

In this formulation, N represents the size of the dataset or the number of observations y_i. Each observation y_i (for $i=1,\ldots,N$) represents a data point of the autoconversion rate and is drawn independently and identically distributed (i.i.d.) from a normal distribution with mean μ and variance σ^2. The mean μ itself follows a normal prior distribution with mean γ and variance $\sigma^2 v^{-1}$, while the variance σ^2 follows an inverse-Gamma prior distribution with shape parameter α and scale parameter β.

As a result, they obtain a joint prior distribution over μ and σ^2 known as Normal-Inverse-Gamma, characterized by the parameters γ, v, α, and β, with $\gamma \in \mathbb{R}$, $v>0$, $\alpha>1$, $\beta>0$. To elaborate further, γ serves as the prior means for μ, representing our initial belief about the data's central tendency before observing any samples. v is related to the measure of the strength or weight of the prior belief μ, with a larger v indicating more confidence in our prior mean γ and a smaller v allowing for more uncertainty. The variance of μ is given by $\sigma^2 v^{-1}$, making v a scaling factor for our trust in the prior mean. α is the shape parameter of the Inverse-Gamma distribution for σ^2, influencing the shape of the variance distribution. Larger α values lead to a distribution that is more peaked (narrower around its mean), implying more certainty about σ^2. On the other hand, β is the scale parameter of the Inverse-Gamma distribution for σ^2, affecting the spread of the variance distribution. A higher β allows the distribution to accommodate larger values for σ^2, which means it can capture a wider range of possible variances and allows for greater variability. These parameters work together to capture uncertainties in both the mean and variance of the data.

https://doi.org/10.1017/eds.2024.37 Published online by Cambridge University Press

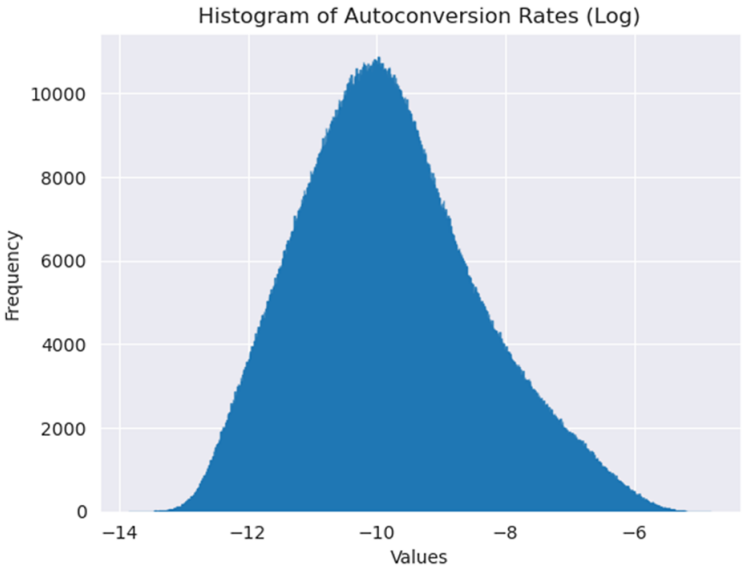

Figure 3. *Histogram of the log-transformed autoconversion rates.*

As shown in Equation 2.4, the target values $y_1, \ldots, y_N$, which in our case are the autoconversion rates, are assumed to be normally distributed (Gaussian). We acknowledge that the distribution of our observed data, after normalization, does not perfectly follow a Gaussian distribution. This is due to the inherent complexity of atmospheric simulation data, which aims to capture real-world conditions. Achieving a perfect Gaussian distribution with real-world data is challenging. However, our data distribution remains reasonably close to Gaussian, as shown in Figure 3. To quantify this, we computed the Kullback–Leibler (KL) divergence between our data distribution and a Gaussian distribution. The KL divergence value of 0.02 indicates a relatively small difference, suggesting that while our data distribution is not perfectly Gaussian, it is sufficiently similar for the purposes of our analysis. KL divergence is a measure of how one probability distribution diverges from another, with lower values indicating greater similarity. In this context, a KL divergence of 0.02 suggests that the Gaussian assumption is reasonable for our data, even though it is not a perfect fit. We have included a discussion of these findings and their implications in Section 3.5.

Therefore, with the Normal-Inverse-Gamma distribution serving as a prior, the marginal likelihood or model evidence is established as the probability of an observation, y_i, given the parameters of the prior, denoted as $\boldsymbol{m}$, where $\boldsymbol{m} = (\gamma, v, \alpha, \beta)$ (Amini et al., 2020). In probability theory, this computation involves marginalizing over the likelihood parameters $\boldsymbol{\theta}$, where $\boldsymbol{\theta} = (\mu, \sigma^2)$, as follows:

$$p(y_i|\boldsymbol{m}) = \int_{\boldsymbol{\theta}} p(y_i|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{m}) \mathrm{d}\boldsymbol{\theta} \tag{2.5}$$

$$= \int_{\sigma^2=0}^{\infty} \int_{\mu=-\infty}^{\infty} p\left(y_i|\mu, \sigma^2\right) p\left(\mu, \sigma^2|\boldsymbol{m}\right) \mathrm{d}\mu \mathrm{d}\sigma^2 \tag{2.6}$$

As shown by Equation 2.6, computing the model evidence (marginal likelihood) is generally non-trivial due to the challenging nature of the double integral, which involves potentially high-dimensional integration over continuous variables. Nonetheless, in certain cases, such as when using conjugate priors (where the posterior distribution belongs to the same family as the prior), analytical solutions are possible.

The Normal-Inverse-Gamma (NIG) prior is conjugate to the normal likelihood. This means that when multiplying the Gaussian likelihood for μ with the Gaussian prior and an inverse-Gamma distribution for σ^2, the resulting posterior distribution remains in the same family as the prior. Put simply, the posterior

https://doi.org/10.1017/eds.2024.37 Published online by Cambridge University Press

distribution of μ and σ^2 after observing the data will also follow a NIG distribution. In the case of NIG priors, the marginal likelihood $p(y_i|\boldsymbol{m})$ has a closed-form expression and is known to follow a Student's t-distribution with 2α degrees of freedom:

$$p(y_i|\boldsymbol{m}) = \mathrm{St}\left(y_i;\gamma,\frac{\beta(1+\nu)}{\nu\alpha},2\alpha\right). \tag{2.7}$$

Given a set of $(\boldsymbol{x_i},y_i)$ pairs, where $\boldsymbol{m_i} = (\gamma_i,\nu_i,\alpha_i,\beta_i) = f(\boldsymbol{x_i};\boldsymbol{w})$, the overall loss function $\mathcal{L}(\boldsymbol{w};\boldsymbol{x_i},y_i)$ for the NN, as defined by Amini et al. (2020), where $\boldsymbol{w}$ represents a set of weights, is formulated as:

$$\mathcal{L}(\boldsymbol{w};\boldsymbol{x_i},y_i) = \mathcal{L}^{\mathrm{NLL}}(\boldsymbol{w};\boldsymbol{x_i},y_i) + \lambda\mathcal{L}^{\mathrm{R}}(\boldsymbol{w};\boldsymbol{x_i},y_i), \tag{2.8}$$

where $\mathcal{L}^{\mathrm{NLL}}$ represents the negative log-likelihood to maximize the model fit and $\mathcal{L}^{\mathrm{R}}$ denotes the regularization term, which is scaled by a regularization coefficient, λ. The NLL term is defined as:

$$\mathcal{L}^{\mathrm{NLL}}(\boldsymbol{w};\boldsymbol{x_i},y_i) = \frac{1}{2}\log\left(\frac{\pi}{\nu_i}\right) - \alpha_i\log\Omega + \left(\alpha_i+\frac{1}{2}\right)\log\left((y_i-\gamma_i)^2\nu_i+\Omega\right) + \log\left(\frac{\Gamma(\alpha_i)}{\Gamma\left(\alpha_i+\frac{1}{2}\right)}\right), \tag{2.9}$$

where $\Omega = 2\beta_i(1+\nu_i)$ and $\Gamma(.)$ represents the gamma function.

The NLL term focuses on fitting the model to the observed data $(\boldsymbol{x_i},y_i)$. If the regularization term $\mathcal{L}^{\mathrm{R}}(\boldsymbol{w};\boldsymbol{x_i},y_i)$ is given a low weight (i.e., λ is small), the model heavily prioritizes minimizing $\mathcal{L}^{\mathrm{NLL}}(\boldsymbol{w};\boldsymbol{x_i},y_i)$. This can lead to overfitting, where the model fits the training data very closely, potentially capturing noise as if it were a true signal. As a result, the model becomes overconfident in its predictions, as it assumes the patterns in the training data will perfectly generalize to unseen data, which is often not the case. The regularization term introduces a form of constraint or penalty to the loss function. By increasing the value of λ, the impact of the regularization term $\mathcal{L}^{\mathrm{R}}(\boldsymbol{w};\boldsymbol{x_i},y_i)$ becomes more significant. This helps to prevent overfitting by discouraging the model from becoming too complex and sensitive to the training data. Instead, the model remains more robust and generalizable, which translates to better handling of uncertainty in predictions. When λ is set too high, the model may become too simple, failing to capture relevant patterns in the data and thus exhibiting excessive uncertainty in its predictions. The regularization coefficient λ determines the trade-off between these two aspects, impacting the overall performance and reliability of the model. Choosing a low value for λ could lead to overconfidence, whereas opting for a high value might result in excessive uncertainty.

Amini et al. (2020) defined the regularization term $\mathcal{L}^{\mathrm{R}}(\boldsymbol{w};\boldsymbol{x_i},y_i)$ as:

$$\mathcal{L}^{\mathrm{R}}(\boldsymbol{w};\boldsymbol{x_i},y_i) = |y_i-\gamma_i|\cdot\Phi, \tag{2.10}$$

where Φ serves as the total evidence. However, instead of utilizing the L_1 error as proposed by Amini et al. (2020), as shown in Equation 2.10, we employ an adjustment to the residuals by incorporating the width of the Student's *t*-distribution (w_{St}), as suggested by Meinert et al. (2023).

We incorporate the modified version proposed by Meinert et al. (2023) because the regularizer formulation proposed by Amini et al. (2020) is insufficient for finding the marginal likelihood parameters. High noise levels in the data can lead to large residuals, which in turn result in large gradients during optimization. This can negatively affect the training process. To mitigate the impact of these large residuals caused by high noise, Meinert et al. (2023) propose scaling the residual by w_{st}, which represents the aleatoric uncertainty. By normalizing the residual with w_{st}, Meinert et al. (2023) aim to reduce the influence of high noise levels on the gradient magnitudes. This normalization helps to prevent large residuals, particularly those resulting from high noise, from disproportionately affecting the optimization process. Consequently, this adjustment inhibits the convergence for large but insignificant residuals associated with aleatoric uncertainty, allowing the model to focus on reducing more meaningful errors. This approach effectively creates an adaptive learning mechanism, where the model becomes more tolerant of errors in high-uncertainty regions while remaining sensitive to errors in low-uncertainty areas. Consequently, the regularization term can be expressed as:

https://doi.org/10.1017/eds.2024.37 Published online by Cambridge University Press

$$\mathcal{L}^{\mathrm{R}}(\boldsymbol{w};\boldsymbol{x_i},y_i)=\left|\frac{y_i-\gamma_i}{w_{St}}\right|^p\cdot\Phi, \tag{2.11}$$

where the parameter p serves as an additional hyperparameter, determining the magnitude of residual impact within regularization. Additionally, w_{St} is computed as:

$$u_{\mathrm{al}}\equiv w_{\mathrm{St}}=\sqrt{\frac{\beta_i(1+\nu_i)}{\alpha_i\nu_i}}. \tag{2.12}$$

Intuitively, as Meinert et al. (2023) state, the standard deviation of a normal distribution, approximated by w_{st}, can be interpreted as the aleatoric uncertainty (u_{al}) of the data (see Equation 2.12). Conversely, the epistemic uncertainty (u_{ep}) is defined by Meinert et al. (2023) as:

$$u_{\mathrm{ep}}\equiv\frac{\sqrt{\mathrm{Var}[\mu_i]}}{\sqrt{\mathbb{E}[\sigma_i^2]}}=\frac{\sqrt{\frac{\mathbb{E}[\sigma_i^2]}{\nu_i}}}{\sqrt{\mathbb{E}[\sigma_i^2]}}=\frac{1}{\sqrt{\nu_i}}. \tag{2.13}$$

The total evidence, denoted as Φ, is defined as (Meinert and Lavin, 2021; Meinert et al., 2023):

$$\Phi=\nu_i+2\alpha_i. \tag{2.14}$$

As explained by Meinert and Lavin (2021) and Meinert et al. (2023), Equation (2.14) provides a better, more intuitive measure of the total evidence because it directly incorporates both the virtual measurements for the mean and the variance. In the conjugate prior NIG distribution, parameters ν_i and $2\alpha_i$ act as virtual observations that encapsulate our prior beliefs. Therefore, the total evidence is determined by combining these two parameters. For the uncertainty quantification part, we employ Meinert et al. (2023)'s approach to measure the aleatoric and epistemic uncertainties, as shown in Equations 2.12 and 2.13, respectively.

2.4. Evaluation of uncertainty quantification

We evaluate the calibration of the uncertainty estimation through the implementation of the spread-skill plot (Wilks, 2011; Haynes et al., 2023). Additionally, to further verify the reliability and calibration of our uncertainty estimates, we also employ the discard test method (Barnes and Barnes, 2021; Haynes et al., 2023).

2.4.1. Spread-skill plot

The spread-skill plot assesses the calibration of the model's uncertainty estimates across different levels of predicted uncertainty. It compares the predicted spread (uncertainty) with the actual prediction errors to determine how well the uncertainty estimates reflect the true uncertainty in predictions.

Algorithm 1 Spread-Skill Plot.

Input: Set of $\{(\mathbf{x}_i,y_i,\widehat{y}_i,\sigma_i)\}_{i=1}^N$
Output: Spread-Skill plot points $\{(S_k,\mathrm{RMSE}_k)\}_{k=1}^K$
Bin the data into K bins based on σ_i
for $k=1$ to K **do**
    $S_k\leftarrow\frac{1}{n_k}\sum_{i\in\mathrm{bin}k}\sigma_i$
    $\mathrm{RMSE}_k\leftarrow\sqrt{\frac{1}{n_k}\sum_{i\in\mathrm{bin}k}(y_i-\widehat{y}_i)^2}$
end for
Plot S_k vs RMSE_k for each bin k
return $\{(S_k,\mathrm{RMSE}_k)\}_{k=1}^K$

https://doi.org/10.1017/eds.2024.37 Published online by Cambridge University Press

The spread-skill plot algorithm, as depicted in Algorithm 1, aims to assess the calibration of uncertainty estimates generated by a predictive model across varying levels of predicted uncertainty. It begins by taking as input a dataset comprising N instances, each characterized by a feature vector $\mathbf{x}_i \in \mathbb{R}^D$, a true target value $y_i \in \mathbb{R}$, a model-predicted value $\widehat{y}_i$, and an associated uncertainty estimate σ_i. The algorithm proceeds by partitioning these instances into K bins based on their respective uncertainty estimates σ_i.

Within each bin k, the algorithm computes two key metrics: S_k, representing the average predicted uncertainty across instances in the bin, and RMSE_k, which denotes the root mean squared error between the model predictions $\widehat{y}_i$ and the true targets y_i within that bin. Specifically, S_k is computed as the average of the uncertainty estimates σ_i for instance within bin k, while RMSE_k is computed as the square root of the average squared differences between $\widehat{y}_i$ and y_i for those same instances.

After computing S_k and RMSE_k for all K bins, the algorithm generates a spread-skill plot by plotting S_k against RMSE_k for each bin k. This plot visually illustrates the relationship between predicted uncertainty and actual prediction errors across different levels of uncertainty. The spread-skill plot assesses how well the model's uncertainty estimates σ_i align with actual prediction errors. Ideally, points should lie along the line $\mathrm{RMSE}_k = S_k$, indicating well-calibrated uncertainty estimates where predicted uncertainties accurately reflect actual prediction errors. Points below the line suggest underconfidence (overestimated uncertainty), indicating room for improvement in uncertainty estimation. Conversely, points above the line indicate overconfidence (underestimated uncertainty), where actual errors are larger than predicted uncertainties, suggesting potential overestimation of model certainty.

2.4.2. Discard test

The discard test evaluates the reliability of uncertainty estimates by progressively discarding predictions with the highest uncertainty and measuring the accuracy of the remaining predictions. This method checks whether the discarded high-uncertainty predictions indeed correspond to larger errors.

Algorithm 2 Discard Test.

1: **Input**: Set of$\{(\mathbf{x}_i, y_i, \widehat{y}_i, \sigma_i)\}_{i=1}^{N}$
2: **Output**: Discard test plot points$\{(p, \mathrm{RMSE}(p))\}$
3: Sort predictions based on σ_i in descending order
4: $\Delta p \leftarrow$ step size for p (e.g., 0.1)
5: **for** $p = 0$ to 1 in steps of Δp **do**
6: Discard top $p \cdot N$ predictions with highest σ_i
7: $\mathrm{RMSE}(p) \leftarrow \sqrt{\frac{1}{(1-p)N} \sum_{i \in \text{remaining}} (y_i - \widehat{y}_i)^2}$
8: **end for**
9: Plot $\mathrm{RMSE}(p)$ vs. p
10: **return**$\{(p, \mathrm{RMSE}(p))\}$

The discard test algorithm, depicted in Algorithm 2, evaluates the reliability of uncertainty estimates generated by a predictive model. It begins by taking as input a dataset comprising N instances, each characterized by a feature vector $\mathbf{x}_i \in \mathbb{R}^D$, a true target value $y_i \in \mathbb{R}$, a model-predicted value $\widehat{y}_i$, and an associated uncertainty estimate σ_i.

The algorithm sorts these instances based on their uncertainty estimates σ_i in descending order. It then systematically evaluates the model's uncertainty estimates by progressively discarding predictions with the highest uncertainty (σ_i) in steps determined by Δp. For each step p from 0 to 1, where p represents the proportion of predictions to discard, it computes the root mean squared error of the remaining predictions $\mathrm{RMSE}(p)$.

This RMSE measures the accuracy of predictions after discarding the top $p \cdot N$ predictions with the highest uncertainty. The algorithm generates a discard test plot by plotting $\mathrm{RMSE}(p)$ against p, providing a visual representation of how the accuracy of remaining predictions changes with the proportion of

https://doi.org/10.1017/eds.2024.37 Published online by Cambridge University Press

discarded high-uncertainty predictions. A well-calibrated model would show a decreasing trend in RMSE as p increases, indicating that higher uncertainty predictions correspond to larger errors, while deviations from this trend can highlight issues in uncertainty estimation.

2.5. Testing scenarios

Building upon the methodology of our prior work (Novitasari et al., 2024, attached), where we tested against various scenarios, this study adopts similar testing scenarios, as follows:

1. *ICON-LEM Germany*: In this testing scenario, we evaluate the performance of our ML models using the same data that was utilized during its training process. This dataset encompasses cloud effective radius and autoconversion rates corresponding to various points in three-dimensional space, obtained through ICON-LEM simulations conducted specifically over Germany. The testing dataset differs from that employed for model training, specifically featuring data on 2 May 2013, at 1:20 pm. This testing scenario allows us to evaluate the model's capacity for generalization to different data within the same region and day, taking into account significant variations of weather conditions that underwent substantial changes (Heinze et al., 2017). The dataset consists of approximately 1 million data points.
2. *Cloud-top ICON-LEM Germany*: In this second testing scenario, we assess the performance of our ML model by employing the same dataset as in the previous scenario, albeit with a specific focus solely on the cloud-top information within the data, representing satellite-like data. The dataset utilized in this testing scenario consists of pairs of cloud-top autoconversion rates and cloud-top effective radius, corresponding to 2D spatial points. These points represent a specific range of latitude and longitude, each associated with a particular altitude, specifically the cloud-top height. We derive this 2D cloud-top data from the 3D atmospheric simulation model by selecting the feature value at specific latitude and longitude points where the integrated COT exceeds 1. The integration is carried out by summing values vertically from the cloud-top. The dataset consists of approximately 200 thousand data points.
3. *Cloud-top ICON-NWP Holuhraun*: In this final testing scenario, we utilize completely different data that features diverse meteorological conditions compared to the previous scenarios. Specifically, we utilize cloud-top data from ICON-NWP Holuhraun, collected at distinct locations, times, and resolutions, in contrast to the data used in the prior scenarios. Details of this dataset can be found in the Data section (Section 2.1). Similar to the second scenario, the dataset utilized in this testing scenario consists of pairs of cloud-top autoconversion rates and cloud-top effective radius, corresponding to 2D spatial points. These points represent a specific range of latitude and longitude, each associated with a particular altitude, specifically the cloud-top height. In particular, this involves ICON-NWP Holuhraun data on 1 September 2014 at 1 pm and 4 September 2014 at 2 pm. The dataset consists of approximately 1.7 million and 1.5 million data points, respectively. The model's capacity to generate reliable outcomes when applied to new data (unseen data) is crucial across various practical applications. This ensures accurate predictions for data not encountered during training, demonstrating effective adaptability to diverse geographic locations, and showcasing its capability to deliver optimal results under varying meteorological conditions.

3. Experimental results

In this section, we present the outcomes of our experiments, providing a detailed examination of various aspects related to our current study. Initially, we delve into the process of selecting the evidential regularizer (λ), a crucial hyperparameter that significantly influences the performance of our deep evidential regression models, followed by conducting an evaluation of our uncertainty estimation, employing two different metrics to assess its calibration. Subsequently, we explored two distinct evidential ML models we employed—one simpler and the other more complex—providing insights into

https://doi.org/10.1017/eds.2024.37 Published online by Cambridge University Press

their architectures and comparing the outcomes between the two. Moving forward, we examine the results of autoconversion prediction using simulation data, followed by the forecasting of autoconversion rates directly from satellite data. Finally, we showcase the uncertainty of both data and model in both simulation and satellite data, offering a comprehensive overview of the uncertainty estimation inherent in our study.

3.1. Selection of the evidential regularizer and evaluation of uncertainty estimation

In our prior study (Novitasari et al., 2024, attached), we explored diverse ML models for predicting autoconversion rates. Despite this, we observed that the results exhibited no significant variations among the different models. Consequently, in this section, we opt for one of the models with a fewer number of training parameters from the two NN models we intend to develop using evidential regression, as elucidated in the subsequent section.

Our evidential regression models were trained with varying evidential regularizer coefficients λ ranging from 1e-2 to 1e-9. This training was based on a shallow NN with a single layer consisting of 64 neurons. λ is a regularization coefficient to scale the regularizer of the evidential regression loss, as previously explained in Section 2.3. Additionally, following Meinert et al. (2023), the hyperparameter p – which signifies the extent of the residual impact within regularization – was chosen with the value of 2. The calibration of uncertainty estimation was evaluated using a spread-skill plot, and the most well-calibrated coefficient was selected. The results of this experiment are shown in Figure 4.

The best coefficient λ was found to be 1e-6 as it was the closest to the ideal line in the spread-skill plot. The ideal line represents a scenario where the predicted uncertainty matches the actual error of prediction. A value above the diagonal indicates overconfidence in the model's predictions, resulting in an underestimation of uncertainties, while a value below the diagonal represents underconfidence, leading to an overestimation of uncertainties. This alignment signifies a well-calibrated model, wherein the estimated uncertainties closely approximate the actual uncertainties.

The evaluation of the total uncertainty performance, encompassing both aleatoric and epistemic uncertainties, played a pivotal role in our investigation of evidential regression applied to simulation data. In addition to the skill-spread plot, we also employed a discard test to further evaluate the calibration of the uncertainty estimation. More detailed explanations of the spread-skill plot and the discard test methods can be found in Section 2.4.

The discard test, presented in Figure 5 as a supplementary evaluation metric, further validates the model's calibration by revealing a monotonically decreasing trend from left to right. The discard test evaluates the reliability of uncertainty estimates by progressively discarding predictions with the highest

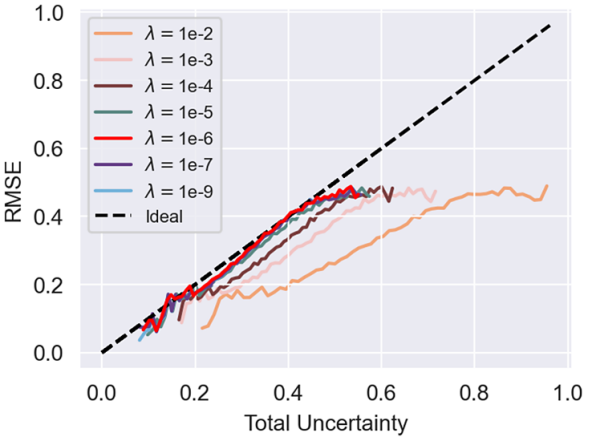

Figure 4. *The spread-skill plot of the deep evidential regression model (based on shallow NN model) with varying evidential regularizer coefficients.*

https://doi.org/10.1017/eds.2024.37 Published online by Cambridge University Press

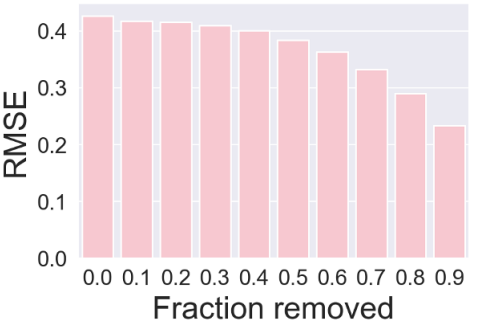

Figure 5. *Evaluation of uncertainty estimation on simulation data (ICON) via discard test.*

Table 1. *Evaluation of autoconversion rate prediction results on simulation data (ICON) using evidential ML models, including both shallow neural network (NN) and deep neural network (DNN) architectures, across various testing scenarios: (1) ICON-LEM Germany, (2) Cloud-top ICON-LEM Germany, and (3) Cloud-top ICON-NWP Holuhraun*

Testing scenario	Model	$\mathbf{R^2}$	MAPE	RMSPE	SSIM	PSNR
1	NN	90.54%	9.14%	11.27%	90.11%	26.30
1	DNN	90.50%	9.19%	11.26%	90.26%	26.29
2	NN	89.88%	10.86%	13.89%	89.96%	25.89
2	DNN	89.95%	10.74%	13.65%	90.18%	25.93
3 (1 Sept)	NN	84.39%	8.56%	10.73%	91.50%	25.81
3 (1 Sept)	DNN	84.99%	8.39%	10.83%	91.77%	25.98
3 (4 Sept)	NN	83.99%	9.26%	11.12%	90.96%	24.79

uncertainty and measuring the accuracy of the remaining predictions. This method checks whether the discarded high-uncertainty predictions indeed correspond to larger errors, confirming a systematic and dependable elimination of low-confidence predictions.

This comprehensive evaluation highlights the evidential regression model's performance in capturing both aleatoric and epistemic uncertainties in simulated data. Both the skill-spread plot and discard test validate the model's robust calibration, offering valuable insights for potential real-world applications.

3.2. Evidential ML models

To facilitate comparison, we train our deep evidential regression models using two distinct neural network architectures: a shallow NN and a DNN, as explained in more detail in Section 2.2. We use a lambda coefficient of 1e-6, determined as the most well-calibrated coefficient based on previous experiments (refer to Section 3.1 for more details), and additionally, following Meinert et al. (2023), the hyperparameter p was chosen with the value of 2.

Table 1 illustrates the comparison between our ML models utilizing shallow NN and DNN, evaluated on ICON-LEM Germany on 2 May 2013 at 1:20 pm. The table also compares these models when focusing on the cloud-top of ICON-LEM Germany at the same timestamp. Additionally, Table 1 showcases a comparison of the ML models on the cloud-top of ICON-NWP Holuhraun data on 1 September 2014 at 1 pm.

The comparison results, as presented in Table 1, demonstrate that all models exhibit comparable performance, delivering quite good results for different testing datasets and scenarios. Despite the fact that the DNN model yields slightly better results in our specific scenario, we consider the NN model (our

https://doi.org/10.1017/eds.2024.37 Published online by Cambridge University Press

current approach) to be preferable since it can achieve comparable outcomes with significantly fewer trainable weights.

3.3. Autoconversion on simulation data (ICON)

This part is very similar to our previous work (Novitasari et al., 2024, attached); however, the difference here is that we conduct the same experiment using a different ML model. In our previous research, we used traditional regression methods and did not include evidential regression. In this continuation work, we apply an evidential regression model. We aim to show that by employing the evidential model, which provides both epistemic and aleatoric uncertainties without additional training or inference, we do not sacrifice the accuracy of the prediction.

We evaluate our NN model using different testing datasets and scenarios associated with the ICON-LEM simulations over Germany and the ICON-NWP simulations over Holuhraun, as explained in Section 2.5.

In summary, those testing scenarios consist of the following: (1) ICON-LEM Germany (different times), (2) Cloud-top ICON-LEM Germany (satellite-like data), and (3) Cloud-top ICON-NWP Holuhraun (different date, time, location, and resolution). Through different testing scenarios, as shown in Table 1, we observe that SSIM values for all cases are approximately around 90%, while R^2 values are also around 90% for the German cases.

SSIM ranges from -1 to 1, where 1 indicates perfect structural similarity, 0 indicates no structural similarity, and -1 would suggest perfect negative structural similarity (which is rare in practice). An SSIM of 90% indicates high structural similarity between our predicted and ground truth images, demonstrating that our model effectively preserves important image features and structures.

R^2 ranges from 0 to 1 and measures the proportion of variance in the dependent variable that is predictable from the independent variable(s). An R^2 of 90% suggests that our model explains 90% of the variability in the data, indicating a strong fit between the predicted and actual values. Higher R^2 values generally indicate better model performance, reflecting that our model captures a significant portion of the data's variability.

Meanwhile, for the Holuhraun cases, despite having a lower R^2 compared to the Germany cases, they still perform reasonably well. In line with our previous research (Novitasari et al., 2024, attached), the observed performance decline is expected since our model was trained solely on ICON-LEM Germany data, rather than ICON-NWP Holuhraun data. Nonetheless, we find that the model trained on ICON-LEM Germany data can still be effectively applied to ICON-NWP Holuhraun data without requiring further re-training, showcasing the model's robustness.

Regarding the MAPE, the results are mostly less than 10% for all cases. This is further supported by the fact that the difference between the prediction results and the ground truth, as depicted in Figure 6 for all cases, are also mostly less than 10%. From the figure, it is evident that despite a slight tendency for overprediction, as indicated by the dominance of the blue color compared to red in the difference parts on the right-hand side, the predictions closely resemble the ground truth.

In general, the results presented in Table 1 and the corresponding Figure 6 demonstrate robust performance. This confirms our approach's ability to accurately estimate autoconversion rates using simulation data. Since the results in this study are similar to those in our previous study [Novitasari et al., 2024, attached], this indicates that we do not sacrifice accuracy by adding the evidential regression component to our model.

3.4. Autoconversion on satellite observation (MODIS)

Similar to Section 3.3, this part of the study closely parallels our previous work (Novitasari et al., 2024, attached). However, while the previous work did not utilize evidential regression, this continuation incorporates it. Our objective is to demonstrate that integrating the evidential model, which provides both epistemic and aleatoric uncertainties without requiring additional training or inference, does not compromise predictive accuracy.

https://doi.org/10.1017/eds.2024.37 Published online by Cambridge University Press

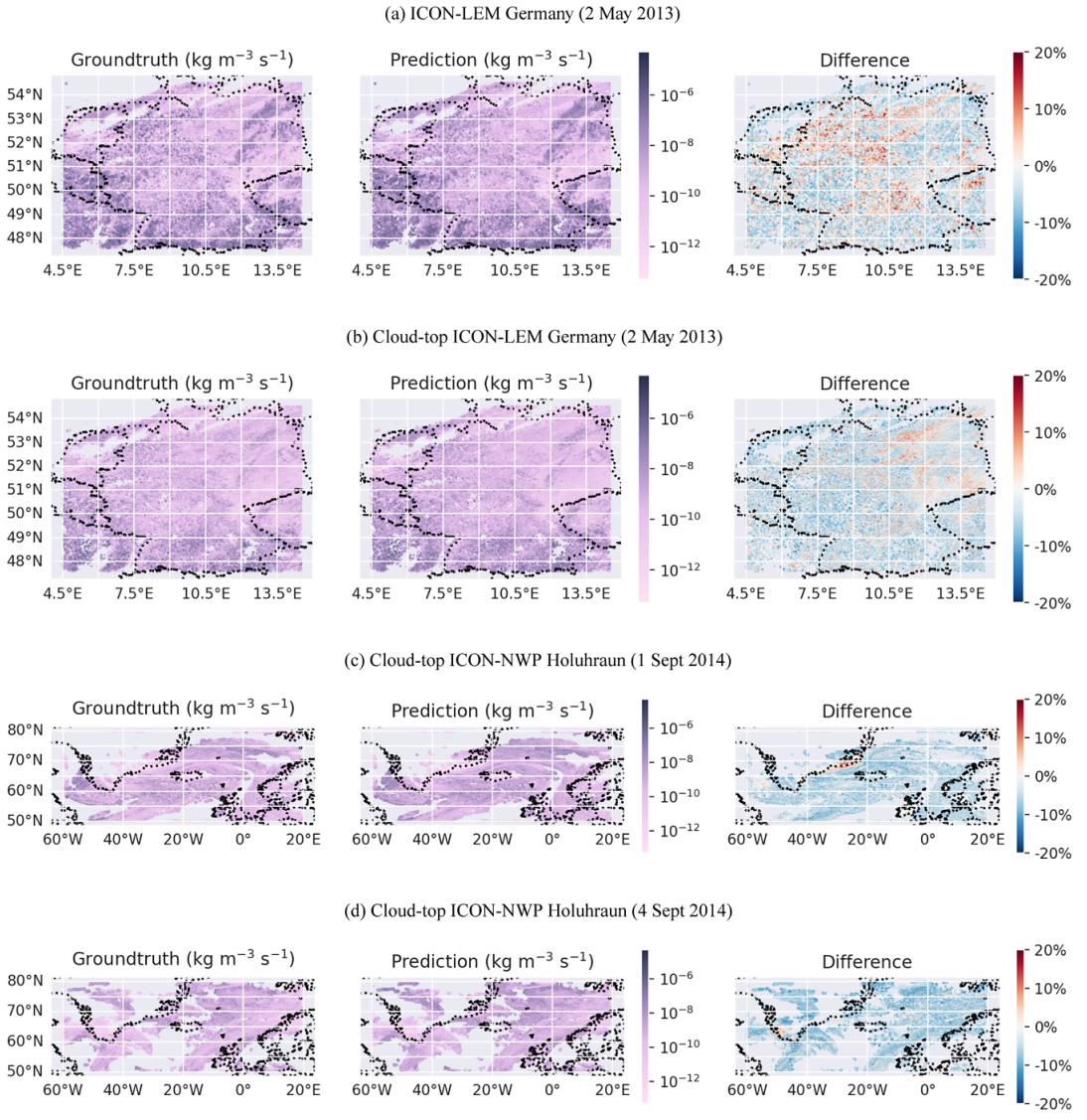

Figure 6. *Visualization of the autoconversion prediction results of ICON-LEM Germany and ICON-NWP Holuhraun. The left side of the image depicts the ground truth, while the middle side shows the prediction results obtained from the NN model. The right side displays the difference between the ground truth and the prediction results. The top image (a) compares the ground truth and predictions from ICON-LEM Germany at a resolution of 1 km, while the second image (b) focuses on cloud-top information only at a resolution of 1 km. The third (c) and fourth (d) figures illustrate the comparison between ground truth and predictions of the ICON-NWP Holuhraun data with a horizontal resolution of 2.5 km, focusing on cloud-top information only.*

This stage involves testing our models on satellite data. Specifically, we utilize MODIS data corresponding to ICON-LEM over Germany and ICON-NWP over the North Atlantic (Holuhraun), aligning with the testing scenarios outlined in Section 2.5 (scenarios 2 and 3). For the Germany case, the latitude range spans from 47.50° to 54.50° N, and the longitude range extends from 5.87° to 10.00° E. This dataset was observed on 2 May 2013, at 13:20 local time. Additionally, for MODIS data associated with

https://doi.org/10.1017/eds.2024.37 Published online by Cambridge University Press

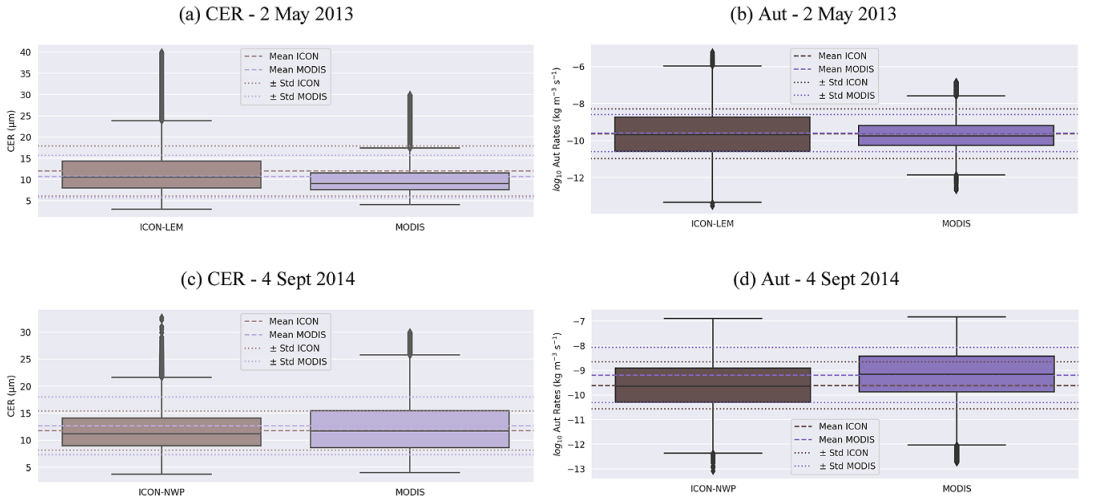

Figure 7. *Mean, standard deviation (Std), median, and percentiles (p25, p75) of cloud-top ICON and MODIS variables over Germany (a and b) and Holuhraun (c and d): cloud effective radius (CER) and autoconversion rates (Aut).*

ICON-NWP Holuhraun, it encompasses latitudes from 60° to 75° N and longitudes from −40° to 0° E. This dataset was captured on 4 September 2014, at 2 pm local time.

Even though satellite predictions cannot be directly compared with simulation results – due to the fact that the ICON-LEM simulation does not put clouds in their exact right places – the autoconversion rates obtained from the simulation output and the predicted autoconversion rates from satellite data demonstrated statistical concordance, for both Germany and Holuhraun cases, as shown in Figure 7. The mean, standard deviation, median, 25th, and 75th percentiles of the autoconversion rates of the cloud-top ICON-LEM Germany and ICON-NWP Holuhraun compared to MODIS are relatively close, especially for the Germany one. Despite the model being trained on ICON-LEM Germany simulation data rather than on ICON-NWP Holuhraun data, the autoconversion rates for the Holuhraun case still indicate statistical closeness between the ICON-NWP Holuhraun simulation and MODIS satellite data predictions. This implies that our approach is capable of estimating autoconversion rates directly from satellite data.

Since the results in this study are similar to those in our previous study (Novitasari et al., 2024, attached), this indicates that we do not sacrifice accuracy by adding the evidential regression component to our model.

3.5. Uncertainty estimation on satellite-like and satellite data

Figure 8 displays the visualizations of aleatoric and epistemic uncertainties on satellite-like data over Germany. Additionally, Figure 9 illustrates aleatoric and epistemic uncertainties, specifically on satellite data over Germany. Aleatoric uncertainties are shown on the left-hand side of the figures, while epistemic uncertainties are shown on the right-hand side of the figures. The brown dot represents the training data points, with the brown line depicting the predictions. Additionally, the purple blob represents both aleatoric and epistemic uncertainties, with darker purple indicating one standard deviation of the uncertainty and the brighter purple showing two standard deviations of the uncertainty. The pink line, available only in simulation data, represents the ground truth.

As mentioned in Section 2.3, our data (normalized autoconversion rates) does not follow a perfect Gaussian distribution. However, it is still reasonably close to Gaussian and can still be used for the purposes of this study. The implications of this are evident in the uncertainty quantification results, where the data is not symmetrically distributed around the modeled uncertainties, as observed in Figures 8 to 11.

https://doi.org/10.1017/eds.2024.37 Published online by Cambridge University Press

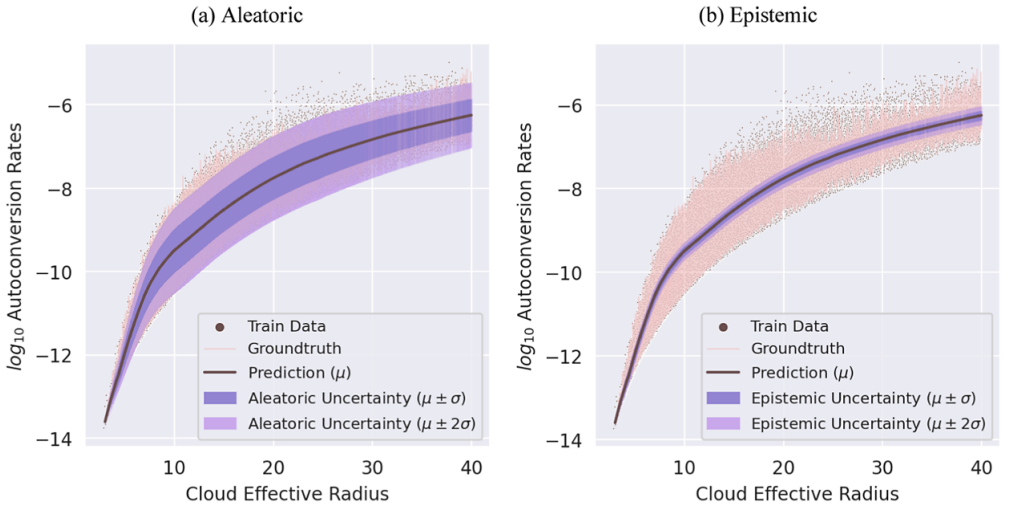

Figure 8. *(a) Aleatoric and (b) epistemic uncertainty estimates of autoconversion rates prediction ($kg.m^{-3}.s^{-1}$) on atmospheric simulation data (ICON) over Germany.*

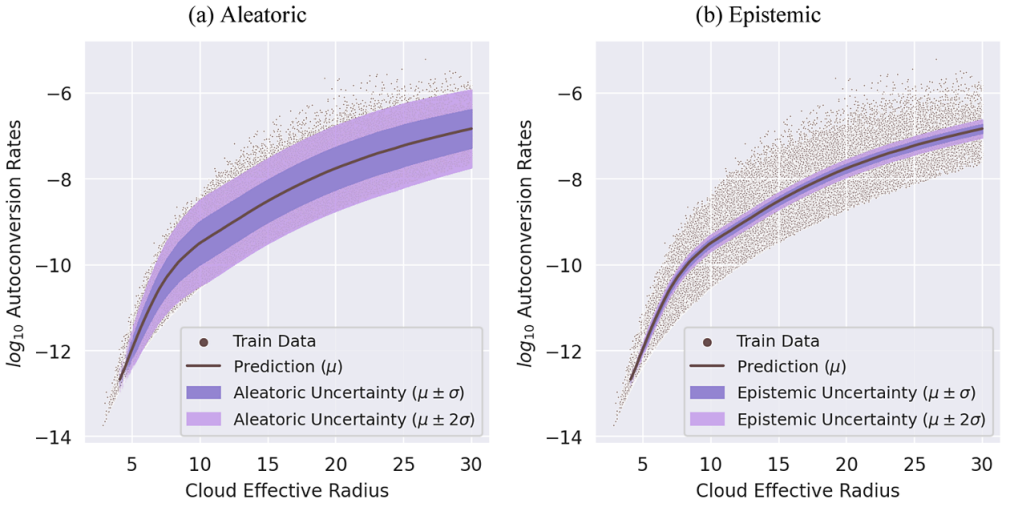

Figure 9. *(a) Aleatoric and (b) epistemic uncertainty estimates of the autoconversion rates prediction ($kg.m^{-3}.s^{-1}$) on satellite data (MODIS) over Germany.*

Based on these results, we infer that the greater the deviation of the data from a Gaussian distribution, the more pronounced the asymmetry in the distribution of the uncertainties.

From these figures, we also see that some data points fall outside the uncertainty bounds. This does not imply that the results are invalid. While the uncertainty quantification is not perfect, it remains a useful tool for interpreting the data. The majority of data points fall within the uncertainty bounds, demonstrating that the uncertainty estimates are generally reliable, even though a small proportion of data points fall outside the expected range. This conclusion is further supported by the evaluation of the uncertainty estimation itself, as explained in Section 3.1.

Turning to the results of the uncertainty quantification itself, these plots reveal a broader range of aleatoric uncertainty compared to epistemic uncertainty. This observation aligns with the fact that a single independent variable can correspond to multiple ranges of dependent variables, as demonstrated in

https://doi.org/10.1017/eds.2024.37 Published online by Cambridge University Press

Figure 8. This observation is further validated by a consistent pattern observed in satellite data, as shown in Figure 9, and by the Holuhraun dataset, which includes both satellite-like and satellite data. Refer to Figure 10 for the satellite-like data and Figure 11 for the satellite data, respectively. These discoveries demonstrate that our results remain aligned and consistent across different datasets.

The results presented show that altering or increasing the complexity of the ML model architecture, using the same dataset, is unlikely to result in a substantial improvement in prediction performance. This finding is further confirmed by our previous experiment in Section 3.2, where adding complexity to our ML model did not significantly enhance the results. The observed lack of improvement is attributed to the dominance of aleatoric uncertainty rather than epistemic uncertainty. Consequently, the focus should shift towards improving data quality, including the addition of more relevant features. Examples of such features include COT per layer, cloud droplet number concentration (CDNC) per layer, and so forth.

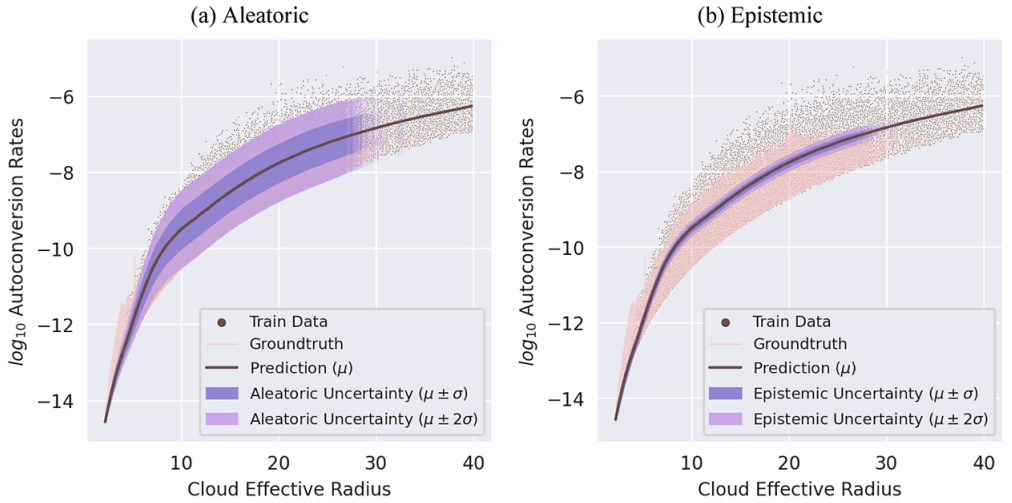

Figure 10. *(a) Aleatoric and (b) epistemic uncertainty estimates of the autoconversion rates prediction ($kg.m^{-3}.s^{-1}$) on atmospheric simulation data (ICON) over Holuhraun.*

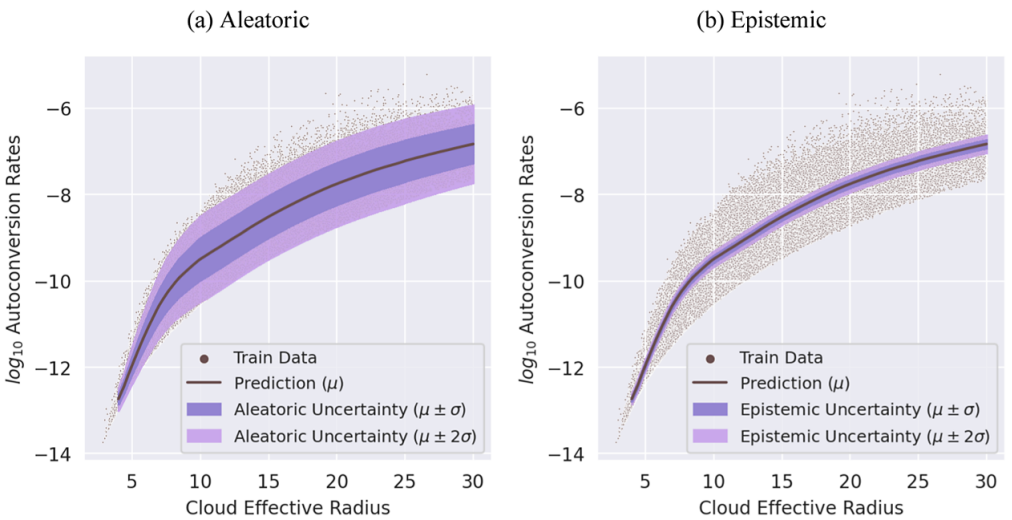

Figure 11. *(a) Aleatoric and (b) epistemic uncertainty estimates of the autoconversion rates prediction ($kg.m^{-3}.s^{-1}$) on satellite data (MODIS) over Holuhraun.*

https://doi.org/10.1017/eds.2024.37 Published online by Cambridge University Press

Unfortunately, these features are currently absent from satellite data. Therefore, future research endeavors should prioritize the implementation of these additional features before integrating them into the ML model. This preliminary step is crucial for laying the foundation for more comprehensive and effective enhancements in prediction results.

4. Concluding remarks

In this study, we provide a computationally efficient solution to unravel the key process of precipitation formation for liquid clouds, the autoconversion process, from satellite data by employing the fewest attributes necessary while still obtaining meaningful results. Furthermore, we employ evidential learning to estimate both data (aleatoric) and model (epistemic) uncertainties, eliminating the need for additional training or inference, hence reducing the overall costs. It has demonstrated good calibration, enhancing the model's credibility, and providing valuable insights for future improvements.

Our findings show that data uncertainty contributes the most to the overall uncertainty, suggesting that modifying the model architecture is unlikely to improve outcomes significantly. This conclusion aligns with both our previous study (Novitasari et al., 2021; Novitasari et al., 2024, attached) and our current study, as explained in Section 3.2, in which we experimented with various ML models to forecast autoconversion rates, nevertheless, the outcomes did not significantly differ across the different models. Instead, in line with other observations by the ML community (Kiureghian and Ditlevsen, 2009; Jain et al., 2020; Singh Sambyal et al., 2022), it would be more effective to prioritize enhancing data quality or incorporating an additional crucial feature, such as COT or CDNC per layer. Unfortunately, satellite data lacks this information, but future efforts will focus on estimating this feature to enhance autoconversion rate estimation and reduce uncertainty.

As the use of ML in atmospheric science gains prominence, the imperative to prioritize trustworthiness becomes crucial. Our method, which leverages evidential regression and is easy to implement in practice, stands as a pivotal contributor to this goal, particularly in refining the reliability of our ML model and comprehensively delving into the intricacies of the uncertainties inherent in the atmospheric science domain. This substantial improvement not only holds the potential to increase the transparency and trustworthiness of ML results but also facilitates a better understanding of the contributing uncertainties. By discerning whether these uncertainties are rooted in data or models, we can strategically tailor our approach. If the uncertainty is data-centered, interventions such as improving data quality can be implemented; if it is model-related, strategies such as refining the ML model or adding samples to the training dataset can be explored. Ultimately, this has the potential to encourage more open discussions and research, especially within the field of atmospheric science.

Open peer review. To view the open peer review materials for this article, please visit http://doi.org/10.1017/eds.2024.37.

Acknowledgments. We sincerely thank and appreciate the editors and reviewers for their valuable feedback and constructive comments, which have significantly contributed to improving the quality of our manuscript. This work used resources of the Deutsches Klimarechenzentrum (DKRZ) granted by its Scientific Steering Committee (WLA) under project ID bb1143.

Data availability statement. The preprocessed data used for training, validation, and testing of this study are available at https://doi.org/10.5281/zenodo.10523401. Additionally, the raw model output data used for the development of the research in the frame of this scientific article is available on request from tape archives at the DKRZ, which will remain accessible for 10 years. The satellite data can be downloaded from the NASA website (https://ladsweb.modaps.eosdis.nasa.gov/search/).

Author contribution. Conceptualization: all authors; Supervision: J.Q., M.R.; Methodology: M.N.; Visualization: M.N.; Writing—original draft: M.N.; Writing—review and editing: all authors. All authors approved the final submitted draft.

Funding statement. This research receives funding from the European Union's Horizon 2020 research and innovation programme under Marie Skłodowska-Curie grant agreement No 860100 (iMIRACLI).

Competing interest. The authors declare no competing interests exist.

Ethical standard. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

References

Albrecht BA (1989) Aerosols, cloud microphysics, and fractional cloudiness. *Science 245*(4923), 1227–1230. http://doi.org/10.1126/science.245.4923.1227.

Amini A, **Schwarting W**, **Soleimany A and Rus D** (2020) Deep evidential regression. In Larochelle H, Ranzato M, Hadsell R, Balcan MF and Lin H (eds.), *Advances in Neural Information Processing Systems*, Vol. *33*. Curran Associates, Inc., pp. 14927–14937. https://proceedings.neurips.cc/paper/2020/file/aab085461de182608ee9f607f3f7d18f-Paper.pdf.

Barnes EA and Barnes RJ (2021) Controlled abstention neural networks for identifying skillful predictions for regression problems. *Journal of Advances in Modeling Earth Systems 13*(12), e2021MS002575. https://doi.org/10.1029/2021MS002575.

Bellouin N, **Quaas J**, **Gryspeerdt E**, **Kinne S**, **Stier P**, **Watson-Parris D**, **Boucher O**, **Carslaw KS**, **Christensen M**, **Daniau A-L**, **Dufresne J-L**, **Feingold G**, **Fiedler S**, **Forster P**, **Gettelman A**, **Haywood JM**, **Lohmann U**, **Malavelle F**, **Mauritsen T**, **McCoy DT**, **Myhre G**, **Mülmenstädt J**, **Neubauer D**, **Possner A**, **Rugenstein M**, **Sato Y**, **Schulz M**, **Schwartz SE**, **Sourdeval O**, **Storelvmo T**, **Toll V**, **Winker D and Stevens B** (2020) Bounding global aerosol radiative forcing of climate change. *Reviews of Geophysics 58*(1). https://doi.org/10.1029/2019RG000660.

Costa-Surós M, **Sourdeval O**, **Acquistapace C**, **Baars H**, **Carbajal Henken C**, **Genz C**, **Hesemann J**, **Jimenez C**, **König M**, **Kretzschmar J**, **Madenach N**, **Meyer CI**, **Schrödner R**, **Seifert P**, **Senf F**, **Brueck M**, **Cioni G**, **Engels JF**, **Fieg K**, **Gorges K**, **Heinze R**, **Siligam PK**, **Burkhardt U**, **Crewell S**, **Hoose C**, **Seifert A**, **Tegen I and Quaas J** (2020) Detection and attribution of aerosol–cloud interactions in large-domain large-eddy simulations with the icosahedral non-hydrostatic model. *Atmospheric Chemistry and Physics 20*(9), 5657–5678. http://doi.org/10.5194/acp-20-5657-2020.

Der Kiureghian A and Ditlevsen O (2009) Aleatory or epistemic? Does it matter? *Structural Safety 31*(2), 105–112. https://doi.org/10.1016/j.strusafe.2008.06.020.

Dipankar A, **Stevens B**, **Heinze R**, **Moseley C**, **Zängl G**, **Giorgetta M and Brdar S** (2015) Large eddy simulation using the general circulation model icon. *Journal of Advances in Modeling Earth Systems 7*(3), 963–986. https://doi.org/10.1002/2015MS000431.

Grosvenor DP, **Sourdeval O**, **Zuidema P**, **Ackerman A**, **Alexandrov MD**, **Bennartz R**, **Boers R**, **Cairns B**, **Chiu C**, **Christensen M**, **Deneke H**, **Diamond M**, **Feingold G**, **Fridlind A**, **Hünerbein A**, **Knist C**, **Kollias P**, **Marshak A**, **McCoy D**, **Merk D**, **Painemal D**, **Rausch J**, **Rosenfeld D**, **Russchenberg H**, **Seifert P**, **Sinclair K**, **Stier P**, **Van Diedenhoven B**, **Wendisch M**, **Werner F**, **Wood R**, **Zhang Z and Quaas J** (2018) Remote sensing of droplet number concentration in warm clouds: A review of the current state of knowledge and perspectives. *Reviews of Geophysics 56*(2), 409–453. http://doi.org/10.1029/2017RG000593.

Gryspeerdt E, **Goren T**, **Sourdeval O**, **Quaas J**, **Mülmenstädt J**, **Dipu S**, **Unglaub C**, **Gettelman A and Christensen M** (2019) Constraining the aerosol influence on cloud liquid water path. *Atmospheric Chemistry and Physics 19*(8), 5331–5347. http://doi.org/10.5194/acp-19-5331-2019.

Haghighatnasab M, **Kretzschmar J**, **Block K and Quaas J** (2022) Impact of holuhraun volcano aerosols on clouds in cloud-system-resolving simulations. *Atmospheric Chemistry and Physics 22*(13), 8457–8472. http://doi.org/10.5194/acp-22-8457-2022.

Haynes K, **Lagerquist R**, **McGraw M**, **Musgrave K and Ebert-Uphoff I** (2023) Creating and evaluating uncertainty estimates with neural networks for environmental-science applications. *Artificial Intelligence for the Earth Systems 2*(2), 220061. https://doi.org/10.1175/AIES-D-22-0061.1.

Heinze R, **Dipankar A**, **Henken CC**, **Moseley C**, **Sourdeval O**, **Trömel S**, **Xie X**, **Adamidis P**, **Ament F**, **Baars H**, **Barthlott C**, **Behrendt A**, **Blahak U**, **Bley S**, **Brdar S**, **Brueck M**, **Crewell S**, **Deneke H**, **Di Girolamo P**, **Evaristo R**, **Fischer J**, **Frank C**, **Friederichs P**, **Göcke T**, **Gorges K**, **Hande L**, **Hanke M**, **Hansen A**, **Hege HC**, **Hoose C**, **Jahns T**, **Kalthoff N**, **Klocke D**, **Kneifel S**, **Knippertz P**, **Kuhn A**, **van Laar T**, **Macke A**, **Maurer V**, **Mayer B**, **Meyer CI**, **Muppa SK**, **Neggers RAJ**, **Orlandi E**, **Pantillon F**, **Pospichal B**, **Röber N**, **Scheck L**, **Seifert A**, **Seifert P**, **Senf F**, **Siligam P**, **Simmer C**, **Steinke S**, **Stevens B**, **Wapler K**, **Weniger M**, **Wulfmeyer V**, **Zängl G**, **Zhang D and Quaas J** (2017) Large-eddy simulations over Germany using ICON: A comprehensive evaluation. *Quarterly Journal of the Royal Meteorological Society 143*(702), 69–100. http://doi.org/10.1002/qj.2947.

Hüllermeier E and Waegeman W (2021) Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning 110*, 457–506.

IPCC (2021) *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* Masson-Delmotte V, Zhai P, Pirani A, Connors SL, Péan C, Berger S, Caud N and Chen Y. Cambridge Univ. Press, (In Press), p. 3949. https://www.ipcc.ch/report/ar6/wg1/downloads/report/IPCC{_}AR6{_}WGI{_}Full{_}Report.pdf.

Jain A, **Patel H**, **Nagalapatti L**, **Gupta N**, **Mehta S**, **Guttula S**, **Mujumdar S**, **Afzal S**, **Mittal RS and Munigala V** (2020) Overview and importance of data quality for machine learning tasks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD'20*. New York, NY: Association for Computing Machinery, pp. 3561–3562. http://doi.org/10.1145/3394486.3406477.

Kendall A and Gal Y (2017) What uncertainties do we need in Bayesian deep learning for computer vision? In Guyon I, von Luxburg U, Bengio S, Wallach HM, Fergus R, Vishwanathan SVN and Garnett R (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA*. Red Hook, NY: Curran Associates, pp. 5574–5584.

https://doi.org/10.1017/eds.2024.37 Published online by Cambridge University Press

Kolzenburg S, **Giordano D**, **Thordarson T**, **Höskuldsson A and Dingwell DB** (2017) The rheological evolution of the 2014/2015 eruption at holuhraun, Central Iceland. *Bulletin of Volcanology 79*(6), 1–16.

Lakshminarayanan B, **Pritzel A and Blundell C** (2017) Simple and scalable predictive uncertainty estimation using deep ensembles. In Guyon I, Von Luxburg U, Bengio S, Wallach H, Fergus R, Vishwanathan S and Garnett R (eds.), *Advances in Neural Information Processing Systems*, Vol. *30*. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf.

Malinin A (2019) *Uncertainty estimation in deep learning with application to spoken language assessment*. PhD thesis, University of Cambridge.

Meinert N, **Gawlikowski J and Lavin A** (2023) The unreasonable effectiveness of deep evidential regression. *Proceedings of the AAAI Conference on Artificial Intelligence 37*(8), 9134–9142. http://doi.org/10.1609/aaai.v37i8.26096.

Meinert N and Lavin A (2021) Multivariate deep evidential regression. Preprint, arXiv:2104.06135.

Novitasari MC, **Quaas J and Rodrigues M** (2021) Leveraging machine learning to predict the autoconversion rates from satellite data. In *NeurIPS 2021 Workshop on Tackling Climate Change with Machine Learning*. https://www.climatechange.ai/papers/neurips2021/59.

Novitasari MC, **Quaas J and Rodrigues MRD** (2024) Cloudy with a chance of precision: Satellite's autoconversion rates forecasting powered by machine learning. *Environmental Data Science 3*, e23. http://doi.org/10.1017/eds.2024.24.

O'Malley T, **Bursztein E**, **Long J**, **Chollet F**, **Jin H**, **Invernizzi L**, et al. (2019) Keras Tuner. https://github.com/keras-team/keras-tuner.

Platnick S, **Meyer KG**, **King MD**, **Wind G**, **Amarasinghe N**, **Marchant B**, **Arnold GT**, **Zhang Z**, **Hubanks PA**, **Holz RE**, **Yang P**, **Ridgway WL and Riedi J** (2017) The MODIS cloud optical and microphysical products: Collection 6 updates and examples from Terra and Aqua. *IEEE Transactions on Geoscience and Remote Sensing 55*(1), 502–525. http://doi.org/10.1109/TGRS.2016.2610522.

Platnick S, **Meyer KG**, **King MD**, **Wind G**, **Amarasinghe N**, **Marchant B**, **Arnold GT**, **Zhang Z**, **Hubanks PA**, **Ridgway B and Riedi J** (2018) MODIS Cloud Optical Properties: User Guide for the Collection 6/6.1 Level-2 MOD06/MYD06 Product and Associated Level-3 Datasets. Available at https://modis-atmos.gsfc.nasa.gov/sites/default/files/ModAtmo/MODISCloudOpticalPropertyUserGuideFinal{_}v1.1{_}1.pdf.

Sambyal AS, **Krishnan NC and Bathula DR** (2022). Towards reducing aleatoric uncertainty for medical imaging tasks. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. IEEE, pp. 1–4. http://doi.org/10.1109/ISBI52829.2022.9761638.

Seifert A and Beheng KD (2006) A two-moment cloud microphysics parameterization for mixed-phase clouds. Part 1: Model description. *Meteorology and Atmospheric Physics 92*(1–2), 45–66. http://doi.org/10.1007/s00703-005-0112-4.

Twomey S (1974) Pollution and the planetary albedo. *Atmospheric Environment (1967) 8*(12), 1251–1256. https://doi.org/10.1016/0004-6981(74)90004-3.

Wilks DS (2011) On the reliability of the rank histogram. *Monthly Weather Review 139*, 311–316.

Zängl G, **Reinert D**, **Rípodas P and Baldauf M** (2015) The ICON (ICOsahedral non-hydrostatic) modelling framework of DWD and MPI-M: Description of the non-hydrostatic dynamical core. *Quarterly Journal of the Royal Meteorological Society 141*(687), 563–579. http://doi.org/10.1002/qj.2378.

Cite this article: Novitasari MC, Quaas J and Rodrigues MRD (2024). Cloudy with a chance of uncertainty: autoconversion rates forecasting via evidential regression from satellite data. *Environmental Data Science*, 3: e45. doi:10.1017/eds.2024.37