Evidential Deep Learning: Enhancing Predictive Uncertainty Estimation for Earth System Science Applications

JOHN S. SCHRECK,[a] DAVID JOHN GAGNE II,[a] CHARLIE BECKER,[a] WILLIAM E. CHAPMAN,[a] KIM ELMORE,[b] DA FAN,[h] GABRIELLE GANTOS,[a] ELIOT KIM,[a] DHAMMA KIMPARA,[c] THOMAS MARTIN,[d] MARIA J. MOLINA,[e,a] VANESSA M. PRZYBYLO,[f] JACOB RADFORD,[g] BELEN SAAVEDRA,[a] JUSTIN WILLSON,[a] AND CHRISTOPHER WIRZ[a]

[a] *NSF National Center for Atmospheric Research, Boulder, Colorado*
[b] *Cooperative Institute for Severe and High-Impact Weather Research and Operations, National Severe Storms Laboratory, Norman, Oklahoma*
[c] *Department of Computer Science, University of Colorado Boulder, Boulder, Colorado*
[d] *University Corporation for Atmospheric Research, Unidata, Boulder, Colorado*
[e] *Department of Atmospheric and Oceanic Science, University of Maryland, College Park, College Park, Maryland*
[f] *University at Albany, State University of New York, Albany, New York*
[g] *Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, Colorado*
[h] *Department of Meteorology and Atmospheric Science, The Pennsylvania State University, University Park, Pennsylvania*

(Manuscript received 13 October 2023, in final form 21 September 2024, accepted 23 October 2024)

ABSTRACT: Robust quantification of predictive uncertainty is a critical addition needed for machine learning applied to weather and climate problems to improve the understanding of what is driving prediction sensitivity. Ensembles of machine learning models provide predictive uncertainty estimates in a conceptually simple way but require multiple models for training and prediction, increasing computational cost and latency. Parametric deep learning can estimate uncertainty with one model by predicting the parameters of a probability distribution but does not account for epistemic uncertainty. Evidential deep learning, a technique that extends parametric deep learning to higher-order distributions, can account for both aleatoric and epistemic uncertainties with one model. This study compares the uncertainty derived from evidential neural networks to that obtained from ensembles. Through applications of the classification of winter precipitation type and regression of surface-layer fluxes, we show evidential deep learning models attaining predictive accuracy rivaling standard methods while robustly quantifying both sources of uncertainty. We evaluate the uncertainty in terms of how well the predictions are calibrated and how well the uncertainty correlates with prediction error. Analyses of uncertainty in the context of the inputs reveal sensitivities to underlying meteorological processes, facilitating interpretation of the models. The conceptual simplicity, interpretability, and computational efficiency of evidential neural networks make them highly extensible, offering a promising approach for reliable and practical uncertainty quantification in Earth system science modeling. To encourage broader adoption of evidential deep learning, we have developed a new Python package, Machine Integration and Learning for Earth Systems (MILES) group Generalized Uncertainty for Earth System Science (GUESS) (MILES-GUESS) (https://github.com/ai2es/miles-guess), that enables users to train and evaluate both evidential and ensemble deep learning.

SIGNIFICANCE STATEMENT: This study demonstrates a new technique, evidential deep learning, for robust and computationally efficient uncertainty quantification in modeling the Earth system. The method integrates probabilistic principles into deep neural networks, enabling the estimation of both aleatoric uncertainty from noisy data and epistemic uncertainty from model limitations using a single model. Our analyses reveal how decomposing these uncertainties provides valuable insights into reliability, accuracy, and model shortcomings. We show that the approach can rival standard methods in classification and regression tasks within atmospheric science while offering practical advantages such as computational efficiency. With further advances, evidential networks have the potential to enhance risk assessment and decision-making across meteorology by improving uncertainty quantification, a longstanding challenge. This work establishes a strong foundation and motivation for the broader adoption of evidential learning, where properly quantifying uncertainties is critical yet lacking.

KEYWORDS: Uncertainty; Classification; Data science; Machine learning; Neural networks; Regression

1. Introduction

Uncertainty is an inherent aspect of any prediction (Abdar et al. 2021), yet it is often overlooked or not communicated effectively alongside the prediction itself (Gneiting et al. 2007). The ability to provide well-calibrated and robust predictive uncertainty estimates can be highly valuable, allowing users to understand the reliability (Rel) of predictions and make more informed decisions based on them (e.g., Nadav-Greenberg and

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/AIES-D-23-0093.s1.

Corresponding author: John S. Schreck, schreck@ucar.edu

DOI: 10.1175/AIES-D-23-0093.1 e230093

Joslyn 2009; Kendall and Gal 2017). For model developers, accurate predictive uncertainty estimates can help identify challenging cases and determine when a model may be operating outside its training domain (Kendall and Gal 2017; Karpatne et al. 2017). Additionally, by connecting uncertainty estimates with other analysis tools, researchers can gain insights into the input sensitivities that influence uncertainty levels, allowing a better understanding of the factors that drive these uncertainties (Herman and Schumacher 2018; Liu et al. 2020). In the realm of machine learning (ML), various approaches have emerged for quantifying total predictive uncertainty. These include well-established methods such as bagging (Breiman 1996), Gaussian processes (Rasmussen and Williams 2006), quantile regression (Koenker 2005), and newer conformal methods (Romano et al. 2019; Stankeviciute et al. 2021; Angelopoulos and Bates 2022). However, it is essential to recognize that existing methods often come with inherent limitations. Conventional techniques, such as bagging, may encounter difficulties in capturing the full spectrum of uncertainty, particularly when dealing with intricate, multimodal distributions. Moreover, many established methods often lack the ability to decompose predictive uncertainty into its underlying components, hindering a deeper understanding of uncertainties. Additionally, handling custom probability distributions can be challenging for some of these methods, limiting their adaptability to specific problem domains.

Traditionally, uncertainty quantification (UQ) within the Earth system science community has been pursued through the development and enhancement of physics-based numerical models, which generate probabilistic forecasts using ensembles of deterministic forecasts that vary in initial conditions, boundary conditions, or model specifications (Leith 1974). One notable advantage of these methods is their strong foundation in the true physics of atmospheric/oceanic motion. However, deterministic numerical model ensembles come with considerable computational costs and often lack proper uncertainty calibration (Vannitsem et al. 2018). To combat the computational expense and lack of calibration, UQ has been attempted through statistical linking functions. Two prominent techniques are ensemble model output statistics (Gneiting et al. 2005), in which a parametric distribution is prescribed and fit, and Bayesian model averaging (Raftery et al. 2005), where the UQ takes the form of a weighted mixture distribution.

The use of modern ML for Earth system UQ has been the focus of much recent research (McGovern et al. 2017; Haynes et al. 2023), especially within the forecast postprocessing community (Haupt et al. 2021; Schulz and Lerch 2022; Vannitsem et al. 2021). Popular techniques include parametric fitting (Ghazvinian et al. 2021; Rasp and Lerch 2018; Chapman et al. 2022; Guillaumin and Zanna 2021; Barnes and Barnes 2021; Foster et al. 2021; Delaunay and Christensen 2022; Gordon and Barnes 2022), quantile-based probabilities transformed to full predictive distributions (Scheuerer et al. 2020), or creating direct approximations of the quantile function via regression based on Bernstein polynomials (Bremnes 2020).

Many users of uncertainty rely on a single metric, such as probability or ensemble spread, but decomposing uncertainty into different components provides valuable insights into its sources and nature (Kendall and Gal 2017) and can help validate how well a model captures the different uncertainty sources. In statistics, the law of total variance decomposes uncertainty into aleatoric, arising from inherent data randomness, and epistemic, arising from model limitations and insufficient training data (Kendall and Gal 2017). High aleatoric uncertainty indicates the lack of a clear relationship between the model inputs and the target and can only be reduced with the addition of more informative input variables (Herman and Schumacher 2018). High epistemic uncertainty can be reduced by accumulating more data in sparse areas of the input space or by reducing the complexity or flexibility of the model. Such distinction aids in assessing model learning capacity, generalization, and guiding hyperparameter optimization (Karpatne et al. 2017). Advancing this field requires techniques that efficiently decompose uncertainty while achieving general predictive reliability.

While parametric probabilistic ML models, such as those that predict the parameters of a categorical or Gaussian probability distribution, can express aleatoric uncertainty (Nix and Weigend 1994), they do not account for epistemic uncertainty (Amini et al. 2020). The predicted variance only accounts for data variance, not model variance. Ensembles of deterministic ML models (Lakshminarayanan et al. 2017) and sampling methods, such as Monte Carlo dropout (Srivastava et al. 2014; Gal and Ghahramani 2016), approximate epistemic uncertainty by deriving their spread from model perturbations, but if their predictions are single labels or values, then they are not accounting for spread (aleatoric uncertainty) in the data. On the other hand, Bayesian neural networks (Neal 2012), which treat every weight as a random variable, can accurately estimate both aleatoric and epistemic uncertainties but are computationally demanding and challenging to implement for complex architectures. Ensembles of parametric probabilistic models can also be used to approximate aleatoric and epistemic uncertainties (Delaunay and Christensen 2022) but only with a sufficiently large ensemble size (Shaker and Hüllermeier 2020).

The recent rise of evidential neural networks (ENNs) (Sensoy et al. 2018; Amini et al. 2020; Ulmer et al. 2021) offers a promising solution that strikes a balance between efficiency and accuracy while providing an effective approach to estimate both sources of uncertainty. ENNs use a single deterministic neural network (NN) while modifying the prediction task to estimate the parameters of a higher-order evidential distribution, which draws relevance from Bayesian data analysis principles (Gelman et al. 2013). This distribution treats the parameters of the target distribution as random variables and models them with an assumed prior distribution (Sensoy et al. 2018; Amini et al. 2020). For multinomial (categorical) and Gaussian distributions, analytical formulations of conjugate prior distributions, such as the Dirichlet and normal inverse gamma (Sensoy et al. 2018; Amini et al. 2020), enable the construction of exact loss functions for NN training. These loss functions consist of a negative log-likelihood component to maximize the fit to the data and a regularizer term to minimize evidence allocated to incorrect predictions and inflate the conditional uncertainty (Sensoy et al. 2018).

In this work, we introduce the concept of evidential deep learning (EDL) to the Earth system science community. EDL

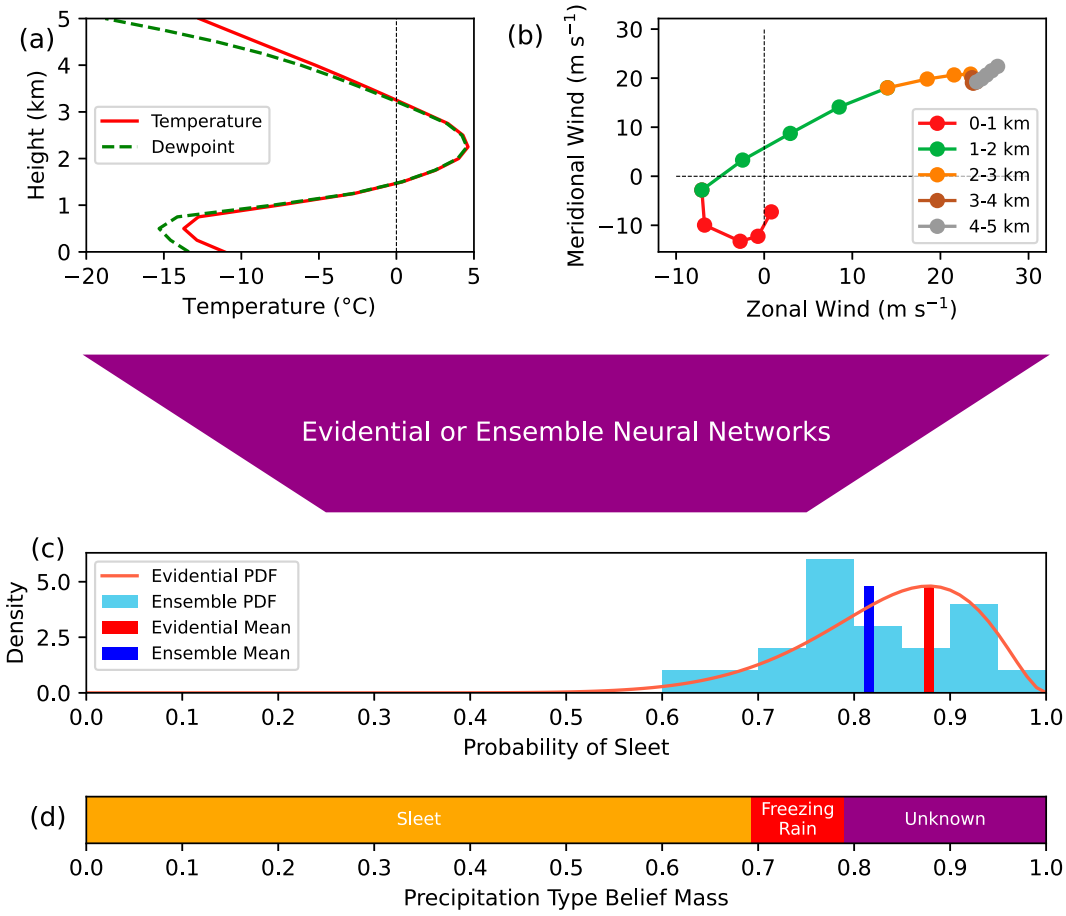

FIG. 1. (a)–(b) Example of temperature, dewpoint, and wind vertical atmospheric profiles, commonly referred to as a sounding, along with (c)–(d) different predictive uncertainty representations for *p* type classification.

represents a relatively recent ML technique known for its ability to provide predictive uncertainty estimates with practical advantages, as discussed in detail by Sensoy et al. (2020). For example, uncertainty estimates are obtained using a single model rather than with an ensemble of models. Our primary focus is on the application of ENNs within the weather and climate domain, where the accurate estimation of uncertainty holds significance in decision-making, as highlighted in previous studies (Shepherd 2009; Bauer et al. 2015). To assess the utility of ENNs, we employ them in both classification and regression tasks. In the classification domain, we showcase an ENN trained to predict winter precipitation type based on simulated atmospheric temperature, dewpoint, and wind vertical atmospheric profiles, commonly referred to as soundings. In regression tasks, our evaluation centers on assessing the model's performance in estimating surface energy fluxes using observed meteorological variables (McCandless et al. 2022).

In addition to practical applications, we have defined three key objectives: 1) to quantitatively assess and compare the predictive skill of evidential versus deterministic neural networks using essential metrics such as Brier skill score (BSS) and root-mean-square error (RMSE). 2) To evaluate the calibration of predicted uncertainties derived from evidential models and ensembles through various analysis techniques. 3) To explore calibration tuning approaches tailored to ENNs, including parameter adjustments such as a loss regularization weight for regression and the fine-tuning of the dropout rate for Monte Carlo (MC) ensembles. These objectives aim to provide a balanced understanding of the potential of evidential architectures in generating uncertainty estimates while maintaining accuracy and emphasize the significance of uncertainties in assessing the limitations of ML models.

2. Problems of interest

a. *Precipitation-type classification*

In the realm of supervised learning, an NN is employed to predict the most likely label or outcome based on a given set of input predictor variables. Figure 1 provides an illustration of model inputs and outputs for a classification problem to identify winter precipitation types. The inputs are shown in Figs. 1a and 1b and are listed in Table 1. They include temperature, dewpoint temperature, zonal, and meridional wind, at equally spaced heights in the atmosphere. The output is a winter

TABLE 1. Summary of the winter precipitation classification and surface layer energy flux regression datasets, including input variables, output variables, and details on the training/validation/testing data splits.

Dataset	Input variables	Output variable (s)	Data splitting
Precip (classification)	Temp (T, °C) Dewpoint temperature (T_{dew}, °C) Zonal wind (U, m s^{-1}) Meridional wind (V, m s^{-1})	Rain Snow Sleet Freezing rain	2015–July 2020: randomly grouped by day, split into training (90%) and validation (10%) sets. Post July 2020 withheld for testing (which is about half the size of the training data).
Surface layer (regression)	Wind speed (10 m; m s^{-1}) Potential temp gradient (10 m to skin; K m^{-1}) Bulk Richardson number (10 m to skin; none) Water vapor mixing ratio gradient (2 m to skin; g kg^{-1}m^{-1})	Friction velocity (m s^{-1}) Sensible heat flux (K m s^{-1}) Latent heat flux (kg m s^{-1})	Randomly split Cabauw dataset from 2003 to 2016 into training (11 years) and validation (2 years) sets, 2015 and 2016 were removed for testing.

precipitation-type label derived from a set of precipitation-type probabilities: $p = (p_{rain}, p_{snow}, p_{sleet}, p_{frzrain})$ (Fig. 1c).

The contrast between the two modeling approaches investigated here (ensemble vs evidential) when predicting the "sleet" label is illustrated in Fig. 1c. In the case of the ENN, the predictions are represented by a smooth curve, offering a continuous and comprehensive quantification of possible outcomes. Conversely, an ensemble approach such as k-fold cross validation (CV), where the training data are divided into k smaller sets, or MC sampling provides discrete point estimates, as seen in the figure, which are obtained by collecting multiple model predictions. It is important to note that these point estimates require an ensemble of a fixed size to approximate the continuous probabilistic insights provided directly by the ENN. This distinction highlights the effectiveness of the ENN in capturing the full spectrum of uncertainty for the full set of labels, as opposed to the more discrete and ensemble-based representation of uncertainty.

The input training data are simulations of atmospheric profiles from the NOAA Rapid Refresh numerical weather prediction model every 250 m from 0 to 5 km above ground level, and the observed outcome is a series of crowd-sourced observations from NOAA's Meteorological Phenomena Identification Near the Ground (mPING) project, occurring within a given volume (13 km) and time frame (1 h) (see Table 1). Within a given volume, we aggregated reports and selected the p type that was the most reported. While this approach captures the dominant precipitation type within a volume, it does not fully account for potential subgrid-scale variability due to meteorological and societal factors influencing the crowd-sourced observations. However, aggregating the most frequently reported precipitation type provides a reasonable representation of the overall conditions for the modeled volume. A small percentage of cases had physically inconsistent precipitation reports, which were filtered out based on criteria involving the wet-bulb temperature and other conditions. Figure 2a shows the multilayer perceptron (MLP) architecture for predicting p type class probabilities.

b. *Surface layer energy flux regression*

This problem aims to train an ML parameterization of the surface layer energy flux (friction velocity, sensible heat, and latent heat) from near-surface atmospheric conditions to replace existing parameterizations in weather and climate models (McCandless et al. 2022; Muñoz-Esparza et al. 2022). Figure 2b depicts the MLP architecture for this regression scenario. The surface layer energy flux problem uses flux tower data from 2003 to 2017 at the KNMI Cabauw site in the Netherlands (Bosveld et al. 2020), which has been used for validating surface layer parameterizations and land surface models since 1972. Details on the inputs and outputs to the model and dataset splitting are provided in Table 1.

3. Methods

In classification problems, ML models output a number associated with each of K known classes for an event. Although predictions cover the volume of an event, the single-point localized observations used for verification may not fully capture the spatial and temporal variability present within a given precipitation/weather event. To model the potential variability in observed outcomes for an event, we can use the multinomial distribution, which generalizes the categorical distribution to the outcomes of m repeated observations under the same event conditions. Formally, given K outcome categories, m observations, and categorical probabilities $\mathbf{p} = \{p_1, \ldots, p_K\}$, the probability of observing the histogram $\mathbf{y} = \{y_1, \ldots, y_K\}$, where y_k is the number of times category k was observed, is

$$p(\mathbf{y}|m, \mathbf{p}, K) = \frac{m!}{\prod_{k=1}^{K} y_k!} \prod_{k=1}^{K} p_k^{y_k}. \quad (1)$$

Given a covariate vector $\mathbf{x}$, we train an ML model to estimate $\mathbf{p}(\mathbf{x})$ by predicting $\hat{\mathbf{p}}(\mathbf{x}) = \{\hat{p}_1(\mathbf{x}), \ldots, \hat{p}_K(\mathbf{x})\}$. We construct a training dataset $\mathscr{D} = \{\mathbf{x}_n, \mathbf{y}_n\}_{n=1}^{N}$, where N is the number of

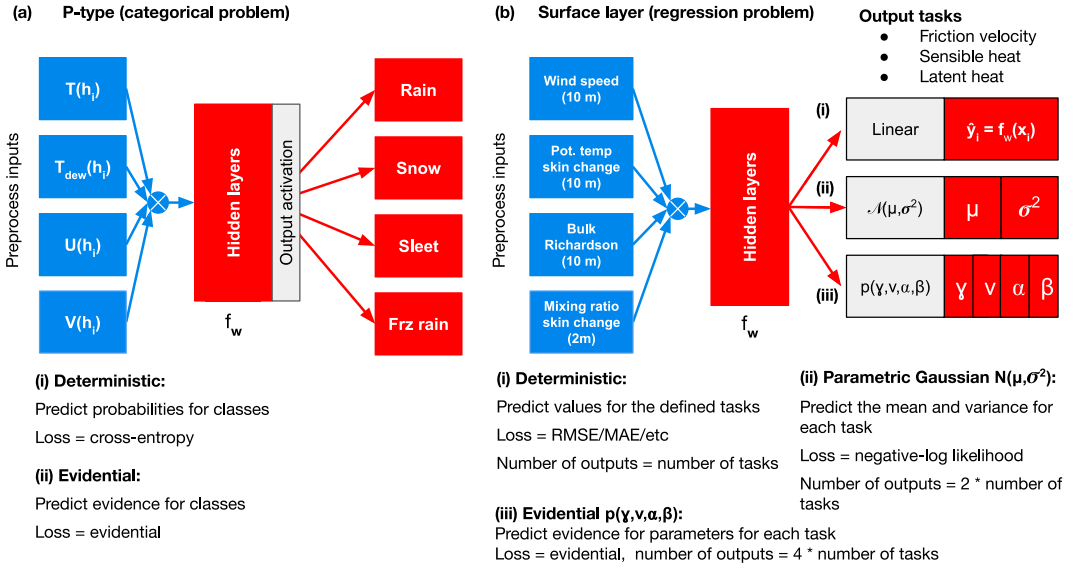

FIG. 2. (a) Deterministic and evidential MLP architectures for predicting class probabilities in the precipitation-type categorical dataset. (b) Architectures for predicting parameters in the surface layer regression dataset, including Gaussian (μ, σ^2) and normal-inverse-gamma (γ, ν, α, β) distributions. In both architecture diagrams, the NN is represented by $f_{\mathbf{w}}$, where $\mathbf{w}$ are the trainable parameters.

data points, that we then input into the ML model repeatedly in order to adjust the parameters (weights) $\mathbf{w}$ of the ML model to maximize their likelihood given the training data. This statistical optimization process, called the maximum likelihood estimation, involves the model ingesting the inputs and predicting outputs y to minimize the negative log-likelihood loss function:

$$\mathscr{L}(\mathbf{w}|\mathscr{D}) = -\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K} y_{n,k}\ln[\hat{p}_{n,k}(\mathbf{x}_n)], \tag{2}$$

where $\mathbf{w}$ are the model parameters, $y_{n,k}$ is a binary indicator for whether the nth sample belongs to class k, and $\hat{p}_{n,k}(\mathbf{x}_n)$ is the predicted probability of the nth sample belonging to class k. Maximum likelihood estimation assumes that we are seeking one fixed set of $\mathbf{w}$ that best explains our data due to its origins in frequentist statistics (MacKay 2003). Thus, these predicted probabilities only account for aleatoric uncertainty from the variation in the outcome for each input and motivate the exploration of other methods to account for epistemic uncertainty.

a. *Model ensembles*

A variety of ensemble methods can be used to obtain uncertainty estimates such as repeating the maximum likelihood estimation process with different random initializations (deep ensembles) or by resampling the training data. Both these approaches can produce ensembles of probabilistic ML models (Lakshminarayanan et al. 2017; Gal and Ghahramani 2016; Dietterich 2000). Another approach is the Bayesian NN (Neal 2012), which treats the weights of the NN as random variables, each represented by a parametric probability distribution. Both ensemble and Bayesian approaches allow for estimating aleatoric and epistemic uncertainties, extending the model's ability to capture and express uncertainty in predictions. While ensembles are straightforward to train, they may require a large number of members or repeated samples to produce robust uncertainty estimates. On the other hand, Bayesian NNs require more weights per model compared to standard NNs and are less likely to converge to results that perform well deterministically and produce robust uncertainty estimates.

We use three ensemble generation methods: CV (k-fold = 20), deep ensembles (ensembles of size 20), and MC dropout ($N_{\mathrm{MC}} = 20$). The ensemble size of 20 was chosen because we observed comparable characterization of uncertainty relative to larger ensembles. For CV, the dataset was divided into folds (subsets), with each fold serving as the validation set for a different model instance. Deep ensembles involved training multiple instances of the same network with different random initializations (Lakshminarayanan et al. 2017), resulting in variations in model performance once training completes. MC dropout randomly deactivated neurons during the training and inference, leading to ensemble predictions (Gal and Ghahramani 2016). A large ensemble size was not required for CV and deep ensembles, so for a fair comparison, N_{MC} was also set to 20. We apply these ensemble techniques to both classification and regression tasks.

b. *Evidential classification*

The recent EDL approaches aim to find a middle ground between ensemble methods and Bayesian NNs (Sensoy et al. 2018; Amini et al. 2020; Meinert and Lavin 2021; Oh and Shin 2022; Meinert et al. 2023). Instead of using fixed weights for

the NN, EDL predicts the parameters for a higher-order posterior distribution. This higher-order distribution describes the space of possible lower-order distributions that could only be partially sampled by ensemble techniques. By taking this approach, EDL tries to combine the benefits of ensembles and Bayesian NNs. In Bayesian inference, prior information is combined with the observed data using Bayes' theorem to update our beliefs and obtain the posterior distribution. Recall that Bayes' theorem states that the posterior probability of a parameter given the data is proportional to the product of the prior probability and the likelihood of the data given the parameter: posterior $\sim$ prior $\times$ likelihood. By including a prior distribution in our inference process, we can define the space of possible models and thus include the epistemic uncertainty. However, in order to make the inference problem tractable, we must pick a prior distribution that can be analytically related to our posterior distribution.

Given these constraints, the best choice of prior for our multinomial/categorical target distribution is the Dirichlet distribution (Murphy 2007; Hoffman et al. 2013). It is described by parameters $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_K)$, which intuitively are "pseudocounts" or the strength of evidence for outcome k. The higher α_k is relative to the other $\boldsymbol{\alpha}$'s, the higher the probability of distributions $\mathbf{p}$, where p_k is relatively large. The probability density function is

$$f(\mathbf{p}|\boldsymbol{\alpha}) = \begin{cases} \dfrac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} p_k^{\alpha_k - 1} & \text{for } \mathbf{p} \in S_K,\ B(\boldsymbol{\alpha}) = \dfrac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^{K} a_k\right)}, \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

where S_K is the set of all K-dimensional vectors $\mathbf{p}$ with entries summing to 1 and each entry greater than or equal to 0. The variable $B(\boldsymbol{\alpha})$ is the K-dimensional multinomial beta function (Kotz et al. 2004). Note the similar form to Eq. (1). The expected probability for the kth outcome is the mean of the Dirichlet posterior distribution:

$$\mathbb{E}[p_k] = \hat{p}_k = \frac{\alpha_k}{S}. \tag{4}$$

Here, $S \equiv \sum_{k=1}^{K} \alpha_k$. The Dirichlet is chosen because it is the conjugate prior to the multinomial, meaning that the posterior when the Dirichlet prior is combined with a multinomial likelihood is also a Dirichlet. This drastically simplifies the updating process during Bayesian inference because each posterior α_k can be computed by summing the prior α_k with each observed or predicted y_k. For example, with the four precipitation types, if our prior is $\boldsymbol{\alpha} = [1, 1, 1, 1]$ and we observe outcomes $y = [10, 5, 2, 3]$, our posterior becomes $\boldsymbol{\alpha} = [11, 6, 3, 4]$, reflecting updated beliefs about each precipitation type based on the evidence.

Sensoy et al. (2018) also motivated the use of the Dirichlet posterior distribution through the Dempster–Shafer theory (DST) of evidence, which is an extension of Bayesian statistics (Dempster 1968) to decision-making with uncertain evidence. Evidence in the DST classification context is represented as a nonnegative number e_k with higher values indicating stronger evidence for a particular outcome. The uninformed evidence prior is that each outcome has evidence of 1, so summing the prior evidence with how much evidence is observed for each outcome from data or a predictive model, we receive a posterior $\alpha_k = e_k + 1$ and the total amount of evidence $S = \sum_{k=1}^{K} (e_k + 1)$. The $\boldsymbol{\alpha}$ values can then be plugged into a Dirichlet distribution to derive probabilities for each outcome.

DST is extended for decision-making in classification problems through subjective logic (Jøsang 2018). Subjective logic, as a formal framework for modeling uncertainty and subjective beliefs, utilizes belief masses to quantify belief strength and enables the nuanced representation of subjective confidence levels. Within this framework, a belief mass b_k is the amount of evidence assigned to a particular outcome, and uncertainty mass u is the amount of evidence not allocated to any outcome or "I do not know,"

$$b_k = \frac{e_k}{S}, \tag{5}$$

$$u = \frac{K}{S}. \tag{6}$$

The variable K is usually defined as the number of outcomes, assuming no prior evidence for any outcome. Belief masses look similar to probabilities but only have to sum to 1 when u is included:

$$u + \sum_{k=1}^{K} b_k = 1. \tag{7}$$

The color bar in Fig. 1d illustrates the belief masses plus u for a precipitation-type ENN. In the integration of DST and subjective logic, the Dirichlet distribution offers a natural way to model belief masses by encoding prior probabilities that reflect the initial expectations or assumptions about the likelihood of different events occurring.

Figure 2a(ii) shows an MLP architecture employing the "evidential" categorical model. The parametric neural network parameterized by $\mathbf{w}$, and given an input sample $\mathbf{x}_n$, it now predicts the evidence vector $\mathbf{e}_n$, represented by $f(\mathbf{x}_n|\mathbf{w})$, hence the name assignment "evidential MLP." Compared to a deterministic classifier, the only architectural difference is the output activation function, which is taken to be ReLU following Sensoy et al. (2018) rather than softmax, to filter negative evidence. Accordingly, the Dirichlet distribution corresponding with this evidence has parameters $\boldsymbol{\alpha}_n = f(\boldsymbol{x}_n|\mathbf{w}) + 1$. The predicted expected probabilities for the classes are then computed as α_k/S.

Sensoy et al. (2018) proposed several loss functions aimed at training NNs to form multinomial opinions or Dirichlet distributions for classification tasks. We focus on the one recommended by Sensoy et al. (2018):

$$\mathscr{L}_n(\mathbf{w}) = \int_{S_K} \|\mathbf{y}_n - \mathbf{p}_n\|^2 \frac{1}{B(\boldsymbol{\alpha}_n)} \prod_{k=1}^{K} p_{n,k}^{\alpha_{n,k} - 1} d\mathbf{p}_n, \tag{8}$$

$$= \sum_{k=1}^{K} (y_{n,k} - \hat{p}_{n,k})^2 + \frac{\hat{p}_{n,k}(1 - \hat{p}_{n,k})}{S + 1}. \tag{9}$$

In this equation, the subscript n refers to the sample index such that $\boldsymbol{\alpha}_n$ represents the full Dirichlet parameter vector $(\alpha_{n,1}, \alpha_{n,2}, \ldots, \alpha_{n,K})$ predicted by the network for the nth sample. Similarly, $\mathbf{y}_n$ refers to a one-hot vector encoding the ground-truth class, $\mathbf{p}_n$ is the corresponding vector of class probabilities induced by the Dirichlet distribution, for sample n, and $\hat{p}_{n,k}$ is as defined in Eq. (4). One arrives at Eq. (9) from Eq. (8) via a standard conjugate prior integration. Without using conjugate pairs, one cannot arrive at a closed form expression as in Eq. (9).

Equation (9) decomposes the mean-squared error and Dirichlet variance. This decomposition allows the network to update its Dirichlet parameters to simultaneously reduce misclassification error and uncertainty during training. It encourages the network to generate higher Dirichlet parameters (more evidence) for correct class labels while avoiding excessive misleading evidence for incorrect classes, prioritizing data fit over variance estimation. Theoretically, it exhibits learned loss attenuation, preventing arbitrarily high evidence masses for unexplainable samples. The overall batch loss is the sum of sample-wise losses.

However, a limited number of counterexamples may lead to an increased overall loss when decreasing the magnitude of evidence, resulting in potentially misleading evidence for incorrect labels. For example, imagine an NN consistently assigning high probabilities to the rain class after encountering numerous instances of rainy weather patterns. Yet, when encountering a few instances of sleet events with similar atmospheric characteristics, reducing the evidence assigned to rain might initially increase the overall loss due to the scarcity of counterexamples.

To address this limitation, a Kullback–Leibler (KL) divergence term is incorporated into the loss function as a regularizer, penalizing divergences from total uncertainty (the uniform distribution) and guiding the network toward a more balanced distribution of probabilities:

$$\mathscr{L}(\mathbf{w}) = \sum_{n=1}^{N} \mathscr{L}_n(\mathbf{w}) + \nu_t \sum_{n=1}^{N} \mathrm{KL}s[D(\mathbf{p}_n|\tilde{\boldsymbol{\alpha}}_n)\|D(\mathbf{p}_n|\mathbf{1})], \quad (10)$$

where ν_t is an annealing coefficient, $D(\mathbf{p}_n|\mathbf{1})$ is a uniform Dirichlet distribution, and $\tilde{\boldsymbol{\alpha}}_n = \mathbf{y}_i + (1 - \mathbf{y}_i) \odot \boldsymbol{\alpha}_n$ represents the Dirichlet parameter vector after removing the nonmisleading evidence from the original $\boldsymbol{\alpha}_n$ for sample n. The KL term prevents early convergence to the uniform distribution for misclassified samples by gradually increasing its impact via ν_t, allowing the network to explore the parameter space. The term reduces to an exact expression given by

$$\mathrm{KL}[D(\mathbf{p}_n|\tilde{\boldsymbol{\alpha}}_n)\|D(\mathbf{p}_n|\mathbf{1})] = \log\left[\frac{\Gamma\left(\sum_{k=1}^{K} \tilde{\alpha}_{n,k}\right)}{\Gamma(K)\prod_{k=1}^{K}\Gamma(\tilde{\alpha}_{n,k})}\right] + \sum_{k=1}^{K}(\tilde{\alpha}_{n,k} - 1) \times \left[\psi(\tilde{\alpha}_{n,k}) - \psi\left(\sum_{j=1}^{K}\tilde{\alpha}_{n,j}\right)\right], \quad (11)$$

where ψ represents the digamma function and Γ is the Gamma function.

c. *Evidential regression*

The concept of an "I don't know" outcome is not explicitly defined for regression tasks with continuous target variables since the range of possible outcomes is unbounded. Subjective logic is primarily designed for handling uncertainty in categorical or discrete scenarios. Amini et al. (2020) approached regression similarly to Sensoy et al. (2018) but adjusted the formulation to handle continuous outcomes by changing the target distribution and loss function. For a regression dataset $\mathscr{D} = \{x_i, y_i\}_{i=1}^{N}$ with i.i.d. targets y drawn from a Gaussian distribution with unknown mean μ and variance σ^2, Amini et al. (2020)'s evidential regression assumes μ is drawn from a Gaussian distribution, while σ^2 follows an inverse-gamma distribution, which is conjugate to a Gaussian, that allows the estimation of epistemic uncertainty, in contrast to traditional regression models that treat μ and σ^2 as fixed and deterministic when using maximum likelihood estimation:

$$\mu_n \sim \mathscr{N}(\gamma_n, \sigma_n^2 \nu_n^{-1}), \quad (12)$$

$$\sigma_n^2 \sim \Gamma^{-1}(\alpha_n, \beta_n), \quad (13)$$

where $\mathbf{m}_n = (\gamma_n, \nu_n, \alpha_n, \beta_n)$, $\gamma_n \in \mathbb{R}$, $\nu_n > 0$, $\alpha_n > 1$, and $\beta_n > 0$. This formulation results in a higher-order distribution referred to as the "evidential distribution," denoted as $p(\mu_n, \sigma_n^2|\mathbf{m})$, which can be represented by a normal-inverse-gamma distribution:

$$p(\mu_n, \sigma_n^2|\gamma_n, \nu_n, \alpha_n, \beta_n) = \frac{\sqrt{\nu_n}}{\sqrt{2\pi\sigma_n^2}}\frac{\beta_n^{\alpha}}{\Gamma(\alpha_n)}\left(\frac{1}{\sigma_n^2}\right)^{\alpha_n+1} \times \exp\left[-\frac{2\beta_n + \nu_n(\mu_n - \gamma_n)^2}{2\sigma_n^2}\right], \quad (14)$$

$$= \mathrm{St}\left[\gamma_n, \frac{\beta_n(1 + \nu_n)}{\nu_n\alpha_n}, 2\alpha_n\right], \quad (15)$$

where St is the Student's t distribution evaluated with location γ_n, scale $\beta_n(1 + \nu_n)/\nu_n\alpha_n$, and $2\alpha_n$ degrees of freedom. For multitask models, there are four parameters for each model target, as is shown in Fig. 2b(iii). By learning the parameters $\mathbf{m}$ through training, regressive ENN models define full distributions over the likelihood parameters (μ_n, σ_n^2), allowing for a comprehensive representation of uncertainty in the model's predictions. The loss used to train a parametric neural network for predicting $\mathbf{m}$ is computed by taking the negative logarithm of Eq. (14):

$$\mathscr{L}_n^{\mathrm{NLL}}(\mathbf{w}) = \frac{1}{2}\log\left(\frac{\pi}{\nu_n}\right) - \alpha_n \log(\Omega_n) + \left(\alpha_n + \frac{1}{2}\right)\log[(y_n - \gamma_n)^2\nu_n + \Omega_n] + \log\left[\frac{\Gamma(\alpha_n)}{\Gamma\left(\alpha_n + \frac{1}{2}\right)}\right], \quad (16)$$

where $\Omega_n = 2\beta_n(1 + \nu_n)$. Following the approach employed by Sensoy et al. (2018), Amini et al. (2020) included an additional regularizer term to suppress evidence (or raise the uncertainty) in support of incorrect predictions:

$$\mathscr{L}_n^R(\mathbf{w}) = |y_n - \gamma_n|(2\nu_n + \alpha_n), \tag{17}$$

where the first term represents the model error, while the second term is proportional to the total evidence accumulated by the learned posterior. The total loss used during training is finally

$$\mathscr{L}_n(\mathbf{w}) = \mathscr{L}_n^{\mathrm{NLL}}(\mathbf{w}) + \lambda\mathscr{L}_n^R(\mathbf{w}), \tag{18}$$

where the parameter λ is selected to best calibrate the model. If λ is too small, the model tends to overfit the data, while overly large values of λ lead to uncertainty overinflation (Soleimany et al. 2021). As we show in the results section, crucially, the evidential regression model's uncertainty estimates are reliant on the tuning of the λ parameter. Calibrating λ requires a separate validation dataset or prior assumptions derived from similar data.

d. Law of total variance

The law of total variance (LoTV; Casella and Berger 2002) states that for two random variables X and Y on the same probability space, the variance of variable Y may be decomposed as

$$\mathrm{Var}(Y) = \mathbb{E}[\mathrm{Var}(Y|X)] + \mathrm{Var}(\mathbb{E}[Y|X]), \tag{19}$$

where $\mathrm{Var}(Y)$ represents the total variance of the random variable Y, $\mathbb{E}[\mathrm{Var}(Y|X)]$ denotes the expected value of the conditional variance of Y given X, and $\mathrm{Var}(\mathbb{E}[Y|X])$ represents the variance of the conditional mean of Y given X. The first term is often referred to as the "aleatoric" uncertainty or "uncertainty in the data," while the latter represents the "epistemic" uncertainty or "uncertainty in the model's predictions."

For the Dirichlet distribution, these quantities can be computed as

$$\underbrace{\mathbb{E}[\mathrm{Var}(y_{n,k}|p_{n,k})]}_{\text{Aleatoric}} = \mathbb{E}[p_{n,k}(1 - p_{n,k})], \tag{20}$$

$$= \frac{\alpha_{n,k}}{S} - \left(\frac{\alpha_{n,k}}{S}\right)^2 - \frac{\frac{\alpha_{n,k}}{S}\left(1 - \frac{\alpha_{n,k}}{S}\right)}{S+1}, \tag{21}$$

$$\underbrace{\mathrm{Var}(\mathbb{E}[y_{n,k}|p_{n,k}])}_{\text{Epistemic}} = \mathrm{Var}(p_{n,k}), \tag{22}$$

$$= \frac{\frac{\alpha_{n,k}}{S}\left(1 - \frac{\alpha_{n,k}}{S}\right)}{S+1}. \tag{23}$$

See section SIII in the online supplemental material for a full derivation. The epistemic uncertainty computed with the LoTV and the quantity u from DST do not have the same form.

Similarly, application of the LoTV to the normal-inverse gamma distribution results in

$$\underbrace{\mathbb{E}[\mu_n]}_{\text{prediction}} = \gamma_n, \tag{24}$$

$$\underbrace{\mathbb{E}[\sigma_n^2]}_{\text{aleatoric}} = \frac{\beta_n}{\alpha_n - 1}, \tag{25}$$

$$\underbrace{\mathrm{Var}(\mu_n)}_{\text{epistemic}} = \frac{\beta_n}{\nu_n(\alpha_n - 1)}. \tag{26}$$

See section SIV in the online supplemental material for the derivation. Note the relationship between the two uncertainty quantities: the epistemic uncertainty is the aleatoric uncertainty divided by parameter ν_n, which is interpreted as a "virtual observation count" that controls how strongly the prior influences the posterior relative to the evidence. As ν_n increases, the prior dominates, so the data must be very persuasive to move the posterior away from the prior prediction γ_n. With lower ν_n, the data readily override the prior, allowing the posterior to diverge from γ_n. So, ν_n modulates the impact of the data versus the prior, rather than literally counting observations.

The LoTV is not limited to stochastic models; it can also be applied to ensembles of probabilistic models. For instance, in the case of a Gaussian parametric model that predicts (μ_n, σ_n^2), an ensemble can be created using various approaches discussed below. Conversely, in a categorical problem, the variance can be estimated directly from the predicted probabilities and the true labels, as shown in Eq. (S5) in the online supplemental material. Given an input $\mathbf{x}_n$ to the ensemble of trained models, the output consists of a list of predicted means and variances. By using Eq. (19), we can compute the aleatoric and epistemic components.

e. Performance evaluation metrics

For classification problems, we use the Brier score (BS) (Brier 1950) to measure the mean-squared difference between the observed labels and predicted probabilities, defined as

$$\mathrm{BS} = \frac{1}{\mathrm{NK}}\sum_{n=1}^{N}\sum_{k=1}^{K}(y_{n,k} - p_{n,k})^2, \tag{27}$$

where lower BS is better, with 0 being a perfect score. The BSS compares the BS to climatology: $\mathrm{BSS} = 1 - (\mathrm{BS}_{\mathrm{forecast}}/\mathrm{BS}_{\mathrm{climatology}})$, where $\mathrm{BS}_{\mathrm{forecast}}$ is the Brier score of the predicted forecast, while $\mathrm{BS}_{\mathrm{climatology}}$ is the Brier score of the climatological forecast, which is the mean-squared difference between the observed frequency of the event and the predicted probability of the event based on climatology (the long-term historical frequency of the event). BSS ranges from $-\infty$ to 1, with 1 indicating a perfect forecast.

For regression problems, we use the RMSE, defined as

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}, \tag{28}$$

where y_i are the true labels and $\hat{y}_i$ are the predictions. Lower RMSE indicates better performance. Both BS and RMSE,

visualized through the attribute diagram, help to gauge how well each model produces an accurate and calibrated mean forecast. Note that these attribute diagrams do not tell us if the model's predicted uncertainty is calibrated.

For that, we follow the methodology outlined by Haynes et al. (2023). For regression problems, we calculate the dependency of the RMSE on the predicted spread (σ). This approach, depicted in the spread–skill diagram, offers a way to normalize model performance against a baseline and provides insights into the relationship between RMSE and forecast spread. A 1–1 relationship in the spread–skill diagram indicates that the model is calibrated according to its uncertainty estimates.

Additionally, we also assess the calibration of uncertainty estimates using the probability integral transform (PIT) plot and the discard fraction plot. For the regression problem, the PIT represents the quantile of the predicted distribution at which the observed value occurs, calculated by approximating the cumulative distribution function of the predicted distribution and evaluating it with the observed value. A uniform PIT histogram indicates a perfectly calibrated model. The PIT deviation (PITD) score quantifies the deviation from uniformity:

$$\mathrm{PITD} = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(\frac{N_m}{N} - \frac{1}{M}\right)^2}. \tag{29}$$

Here, M is the number of bins, N_M is the count of samples in each bin, and N is the total number of samples. Lower PITD is better, with 0 indicating perfect calibration. However, PITD is sensitive to the number of bins, so we evaluate the PITD skill score relative to the worst-case scenario:

$$\text{PITD skill score} = 1 - \frac{\mathrm{PITD}}{\mathrm{PITD}_{\mathrm{worst}}}, \tag{30}$$

where $\mathrm{PITD}_{\mathrm{worst}}$ represents the worst possible PITD score, which assumes that all the forecasts end up in one of the bounding bins of the PIT histogram. Higher PITD skill scores (up to 1) indicate better calibration relative to the worst case.

Finally, for both classifier and regression problems, we investigate discrimination performance using the discard fraction plot, which shows how the model's ability to differentiate between outcomes improves as prediction confidence grows. It involves sorting the data by uncertainty and progressively removing data points with higher uncertainties to assess any improvement in model performance. The discard improvement (DI) score is a measure of performance improvement when certain percentiles of the uncertainty metric are discarded. Higher values indicate a better DI. The discard fraction provides insights into the correlation between prediction error and uncertainty and is valuable for setting model spread thresholds for operational use. During deployment, predictions above this uncertainty threshold can be discarded or considered "out of confidence" to avoid outputting potentially misleading results in uncertain situations. The threshold for model spread can be set manually based on specific application requirements or according to predefined specifications; for instance, a common approach involves using a 95% threshold for acceptance in other applications.

4. Results

a. Winter precipitation type

Figure 3 compares reliability and resolution (Res) between a deterministic and an ENN (the confusion matrix for each model can be found in the supplemental material). Both attribute diagrams show similar calibration trends and reasonable performance, with the most notable deviation being the slight overprediction of sleet at higher probabilities. The ENN showed slightly better results for rain and snow, whereas the deterministic model performed marginally better for sleet and freezing rain. The main takeaway for this example is that evidential models offer comparable probabilistic predictions to deterministic models and can provide deeper insights about the sources of uncertainty.

To assess the quality of the predicted uncertainties, Fig. 4 demonstrates the discard curves using the BS, where zero represents a perfect model. Although the discard test may not be a true calibration metric, the overall negative and higher slopes for the most uncertain data bins indicate reasonably well-calibrated uncertainty for all classes for both model types. It is worth noting that the ENN and the CV ensemble perform comparably, with the ENN performing slightly better in most cases according to the discard improvement score. These well-calibrated uncertainties were achieved without extensive hyperparameter tuning. While hyperparameters such as dropout rate (for the CV ensemble of deterministic MLPs, see Fig. S2) and the loss weight on the KL divergence term (for the evidential model) can be used to fine-tune calibration to some extent, their influence is limited. In fact, for the evidential model, the uncertainties are predominantly well-calibrated through model optimization on the F1 score alone (see section SVI in the online supplemental material for training and optimization details).

In operational use, consider the rain class with a BS threshold of 0.02, where any prediction with a BS value smaller than 0.02 would be accepted. Since the uncertainty was used to order the dataset in Fig. 4, this BS threshold corresponds to a discard threshold (u^*) for DST uncertainty. Any prediction with uncertainty $u > u^*$ would be considered too uncertain. For rain, using this discard threshold based on uncertainty would allow around 90% of the rain examples to be automatically processed without requiring human intervention.

To illustrate the capabilities of the ENN in a real-world scenario, Fig. 5 presents a case study of a severe winter weather event that impacted the central United States on 17 December 2016. The ENN predicts the p type over the locations where the numerical weather prediction model predicts any kind of precipitation, with the model predicting the probability of each precipitation type conditioned on precipitation occurring. All four precipitation types are predicted over the central United States (Fig. 5a), and the various uncertainty estimates (Figs. 5b–d) are also provided. Aleatoric uncertainty (Fig. 5b)

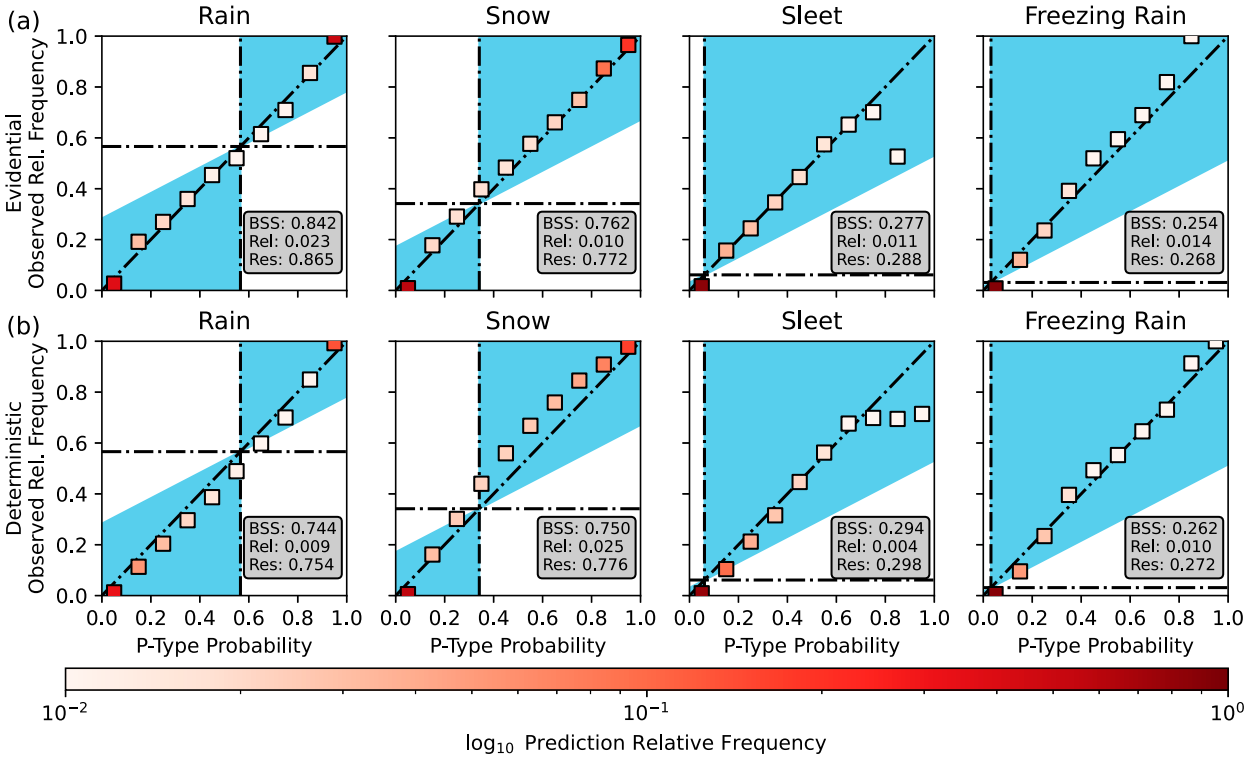

FIG. 3. Attribute diagrams for (a) the evidential model and (b) a deterministic NN model. The columns show the result for each precipitation type. In each subpanel, the diagonal, horizontal, and vertical dashed lines indicate the perfect Rel line, no-Res line, and climatology line. Red-shaded rectangles illustrate the Rel of each model in predicting each class. The legend in each panel displays the BSS, along with the Rel and Res components of the BS divided by the uncertainty component (Murphy 1973). Rel describes the deviation of the predicted probability from the observed relative frequency (lower is better) and Res describes the average difference in the predicted probabilities from climatology, related to sharpness where higher values are better. The blue-shaded area indicates where Res $\geq$ Rel. Points in this region contribute to a positive BSS.

is highlighted primarily where we would expect it, in the transition zones between p types, where soundings could exist that even experienced forecasters would likely have some uncertainty in their assessment of precipitation type. Also notable is the previously noted very strong relationship between DST u (Fig. 5c) and LoTV (Fig. 5d); however, the magnitude of total uncertainty via LoTV is generally much lower. Unlike aleatoric uncertainty, DST u and epistemic uncertainty are most elevated in the center of the freezing rain region and are lower elsewhere, which may be related to a sounding profile that is less frequent in the data.

Last in Fig. 6, we sort the predictions by various uncertainty thresholds and plot composites of the temperature profile to verify that the UQ estimates are physically consistent with meteorological understanding. We expect to see the more certain composites constrained to physically relevant areas of the temperature plane. For example, we would expect the most certain snow composites to have the coldest temperatures throughout the profile and the opposite to be true for rain. For sleet, we would expect there to be a shallower, cooler warm nose (area of the profile above freezing) and a larger near-surface freezing area that would give ample time for the water to refreeze. We used various UQ estimates for sorting [columns (Figs. 6a–c)], which all show expected directional trends with uncertainty.

b. *Surface-layer flux*

In this section, we present a similar evaluation of an evidential regression model trained on the surface layer data from a flux tower to predict energy flux. The key point of comparison, detailed below, is that while the evidential regression model provides a solution for modeling uncertainties like the categorical approach, its calibration is much more sensitive to the weight λ multiplying the KL term in Eq. (18) during training. Furthermore, we observe the regression model infrequently producing unrealistic uncertainties, even when calibrated, potentially posing operational challenges. In addition to the evidential regression results, we also present results from ensembles created using CV and deep ensembles and MC dropout for comparison.

We initially address the issue of model calibration by examining Fig. 7a, which displays the PITD skill score as a function of the KL weight λ for a three-task model that predicts fraction velocity, latent heat, and sensible heat. The dataset's characteristics prevent the model from achieving the maximum PITD skill score at the same λ value for all three tasks,

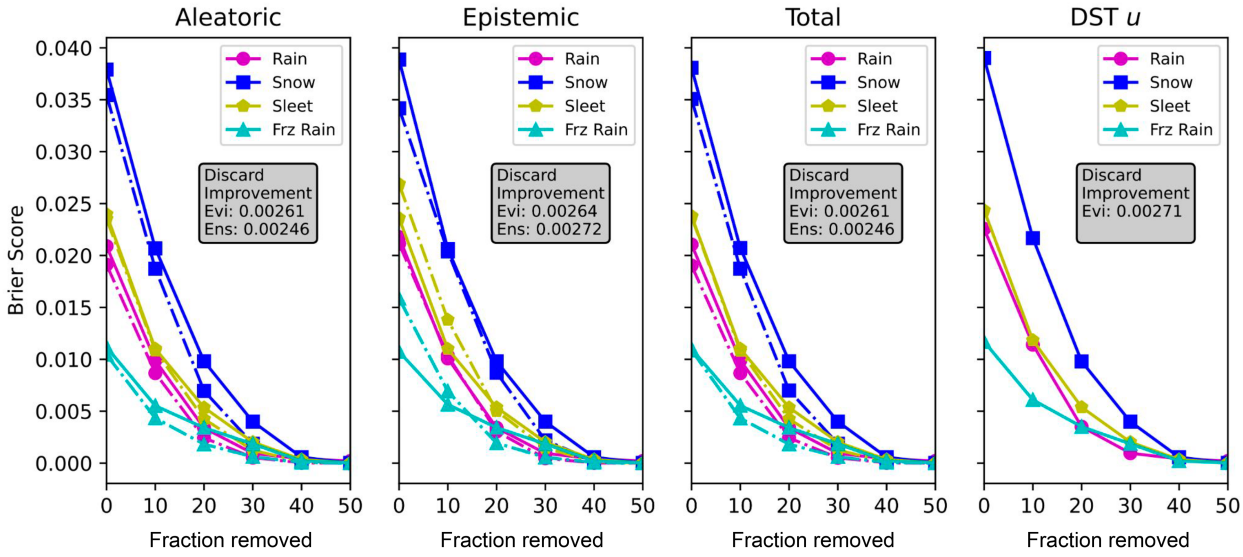

FIG. 4. Discard-test diagrams show the fraction of data points removed from the test set vs the BS computed on the remaining subset. Dashed lines illustrate the ensemble model, while solid lines show the evidential model. The legend in each panel shows the DI score. Solid and dashed lines show the ENN and the CV ensemble results, respectively.

and none of the tasks approached a perfect skill score of one. To further explore this relationship, we analyze Fig. 7b, which depicts the same dependency computed using three single-task models. Effective utilization of the evidential regression model on the SL dataset requires three single-task models to determine the optimal λ values (illustrated in the figure), and furthermore, the values differ relative to the three-task model shown in Fig. 7a. Recall that a flat PIT histogram (PITD equal to zero) does not necessarily guarantee calibration (Chapman et al. 2022; Haynes et al. 2023; Hamill 2001).

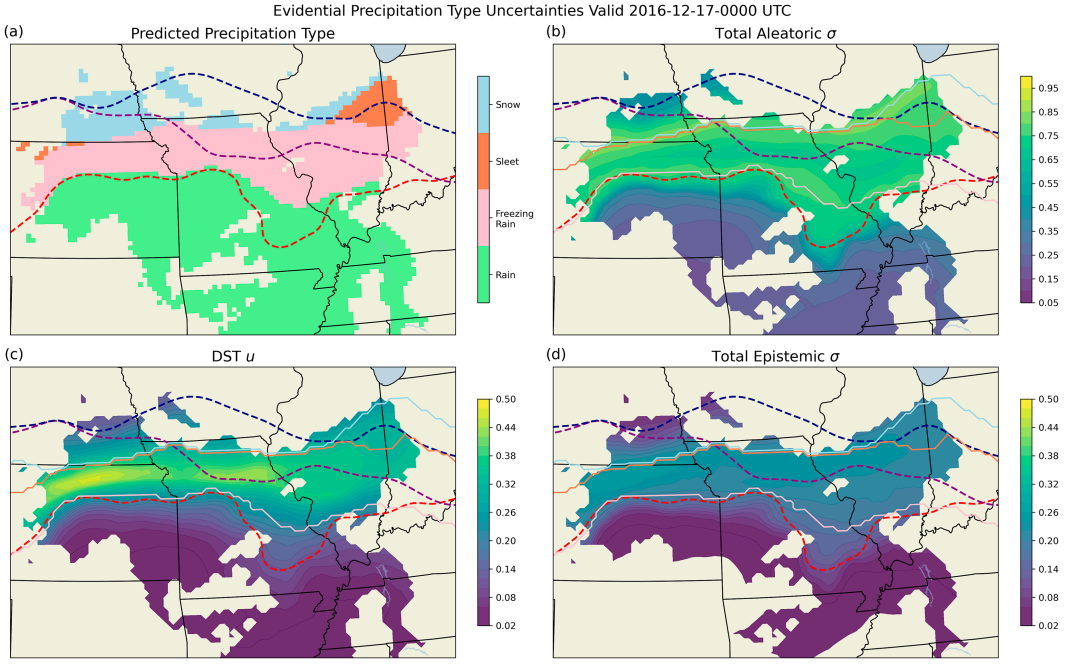

FIG. 5. 17 Dec 2016 precipitation-type predictions and their uncertainties from the evidential model are visualized for the central United States. (a) The most likely precipitation-type prediction and (b)–(d) the total aleatoric, DST, and epistemic uncertainties, respectively. The red, purple, and navy contours indicate the 0°C isotherm at the surface, 2-km AGL, and the 0–5-km AGL maximum, respectively.

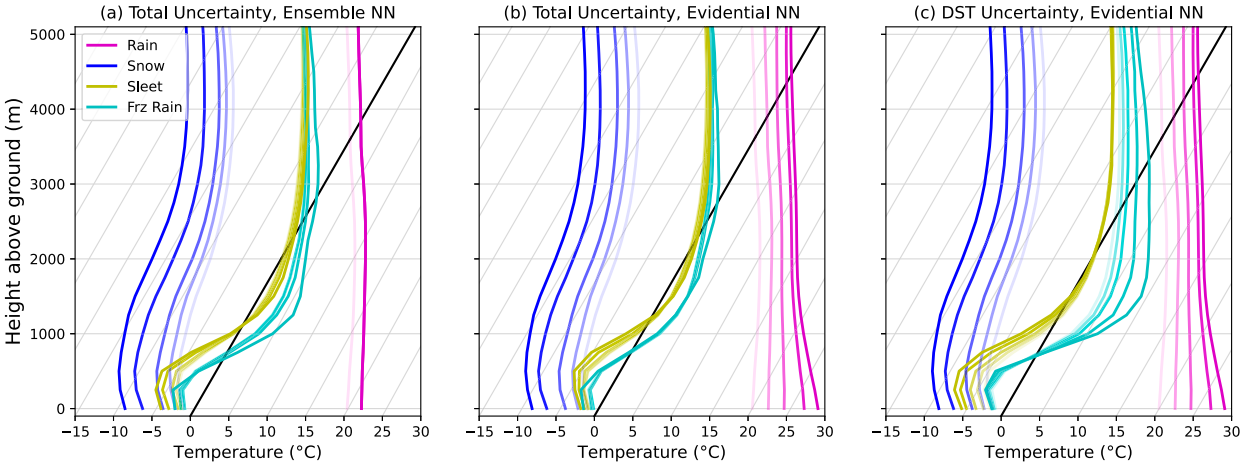

FIG. 6. (a)–(c) Composite soundings are presented lightest-to-darkest using the median of the 10th, 20th, 50th, and 90th percentile most certain predictions generated by the evidential model. The total uncertainty is utilized for (a) the MC-ensemble MLP and (b) the evidential MLP, while the DST uncertainty is employed for (c) the evidential MLP.

With the calibration weights identified, the three models were trained and compared against an ensemble of Gaussian models created using CV splits. Figure 8 displays the reliability diagram for the ENN and the ensemble of Gaussian models (the results for other ensembles are shown in Fig. S5). Overall, both models demonstrate similar reliability and correlation with observations. However, there are some discrepancies between the two. The Gaussian model exhibits less sharpness and correlation for latent heat predictions. The ENN variation shows promise for uncertainty quantification in our regression case study. However, unlike discrete categorical problems, regression tasks face additional calibration challenges due to their unbounded nature. Further research is needed to generalize these findings across diverse regression problems.

How effective is the calibration? Figure 9 shows 2D histograms quantifying the relationship between the computed RMSE for the three model tasks and aleatoric, epistemic, and total uncertainty for both evidential models and Gaussian ensembles. Unlike the precipitation-type problem, the dominant uncertainty for the evidential model is epistemic uncertainty, which contributes the most to the total uncertainty. The 1–1 relationship is observed for epistemic and total uncertainties when both the RMSE and uncertainty are generally small, and this relationship is observed for more than half of the testing data in each case. However, as the quantities increase, the computed RMSE flattens out for sensible and latent heats, while for friction velocity, the relationship appears to be linear with an initial flattening of the RMSE for relatively higher values of uncertainty, which then continues to grow linearly. Note also that none of the models were calibrated according to aleatoric uncertainty; in fact, all of them were underdispersive. Overall, the best PITD skill score and the best fit between the 1–1 line were achieved for friction velocity, which is inherently an easier task to solve.

The Gaussian ensemble showed opposite trends in its uncertainty allocation. For all three outputs, aleatoric uncertainty is favored over epistemic. The RMSE generally increases faster than the aleatoric uncertainty, pointing to the ensemble uncertainty estimate being underdispersive and not accounting for the full range of uncertainty. The evidential model's parametric assumptions about how uncertainty is

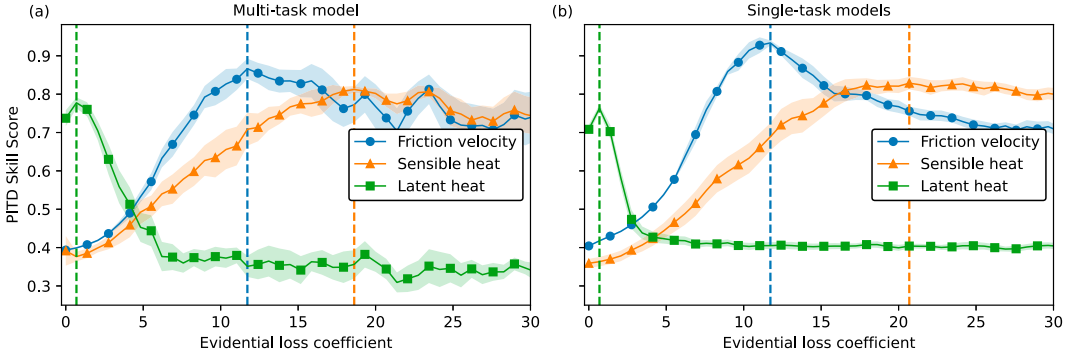

FIG. 7. The PITD skill score as a function of the evidential loss coefficient [λ, Eq. (17)] for (a) one multitask model and (b) three single-task models.

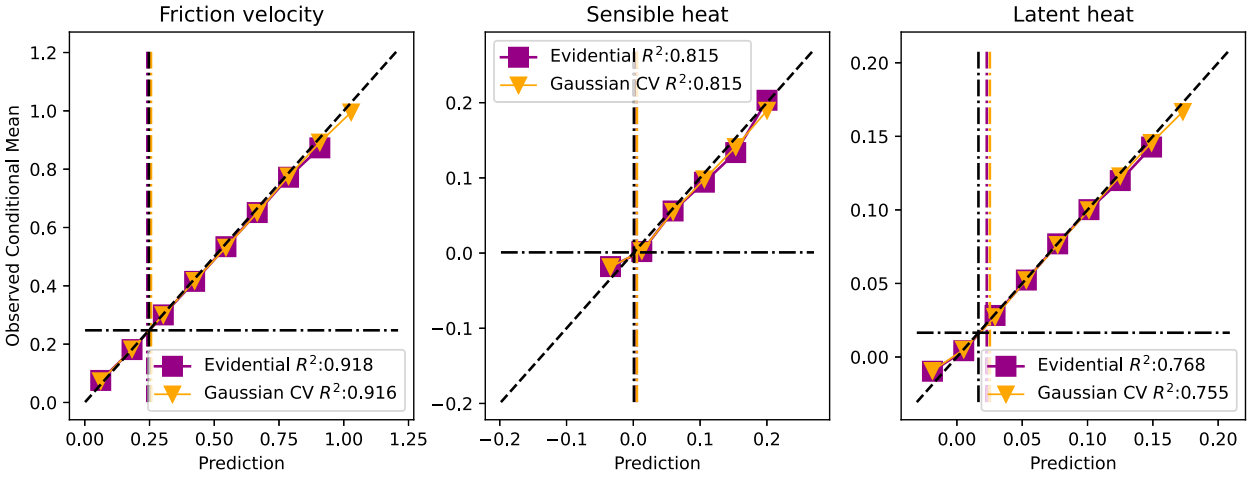

FIG. 8. Rel curves for the cross-validation ensemble using the parametric Gaussian model and the evidential model. The subpanels show the comparison for each model task. In each subpanel, the diagonal, horizontal, and vertical dashed lines indicate the perfect Rel line, no-Res line, and climatology line.

allocated result in a more calibrated total uncertainty estimate for all three outputs.

These results are further corroborated by the PIT histogram and the discard-test, which are shown in Figs. 10a and 10b, respectively. The PIT histogram also shows that friction velocity came the closest to calibration, having the flattest histogram of the three models, followed by latent heat. These quantities also show a hump toward the left middle side of the histogram indicating slight underconfidence by the model. Friction velocity also has a small hump on the right side of the histogram, further indicating some overconfidence on a subset of the data. The sensible heat model, on the other hand, clearly shows a pronounced hump on the left and a smaller one on the right side of the histogram, showing that observations fall near a tail, or fully outside of the predicted distributed, more often than they should. All of these observations are consistent with Fig. 9. All models were subject to extensive hyperparameter optimization, which suggests that there are limitations to the degree of calibration possible on certain datasets when using the evidential regression approach. Section SVI in the online supplemental material details the optimization approach and computational cost, which exhibited considerable variability across models, ranging from around 10 h for a model predicting a single target on graphical processing units (GPUs) to over 75 h for a more complex evidential model predicting multiple targets like friction velocity, sensible, and latent heat fluxes on the same GPU resources.

The discard test shows that for all three models, the predicted aleatoric, epistemic, and total uncertainties are linked to the performance of the model through the RMSE, where the more certain data points have lower RMSE values in each case. This is still the case even though the models possessed different degrees of calibration. Therefore, the uncertainty value can be thresholded such that when the model is in operation, it can be used conservatively when model predictions are too uncertain (e.g., the model will not return a prediction if the predicted uncertainty is larger than the threshold).

While the ENN shows promising performance, caution must be exercised when interpreting the predicted uncertainty values as they may not consistently align with the range of the (in-distribution) training target values (Ovadia et al. 2019). Discrepancies were observed, such as unexpected humps in the epistemic uncertainty distributions at extreme values for certain quantities like friction velocity (Fig. S8). Additionally, different optimized model architectures, despite comparable overall calibration, produced varying uncertainty distributions, especially at the extremes. These findings underscore the need for critical evaluation and validation of the uncertainty estimates to ensure their reliability and alignment with the true nature of the data.

Further analysis revealed distinct diurnal trends in the epistemic uncertainty for friction velocity, sensible heat, and latent heat fluxes. In Fig. 11, the composite epistemic and aleatoric uncertainties vary in concert with the predicted values. During the daytime, flux values increase due to insolation heating the surface and causing convective eddies and evapotranspiration. The epistemic uncertainty increases at a faster rate compared with the aleatoric uncertainty. However, even small changes to the evidential loss weight could change the magnitude of this relationship (Fig. S9), emphasizing the importance of properly calibrating the loss to ensure reliable and physically sensible uncertainty predictions, particularly for latent and sensible heat tasks.

To gain further insight into why the predictions and uncertainties vary with the diurnal cycle, we visualize the relationship between the dominant inputs and the uncertainties, the 10-to-0-m temperature gradient, and the 10-m wind speed (Fig. 12). Each bivariate combination is fed through the ENN. To ensure physical consistency, a bulk Richardson number derived from the chosen wind speed and temperature gradient and a fixed mixing ratio gradient are also provided to the model input vector. Figure 12a shows that variations in sensible heat flux are driven more by the temperature gradient for negative gradients (unstable regime), especially at high wind

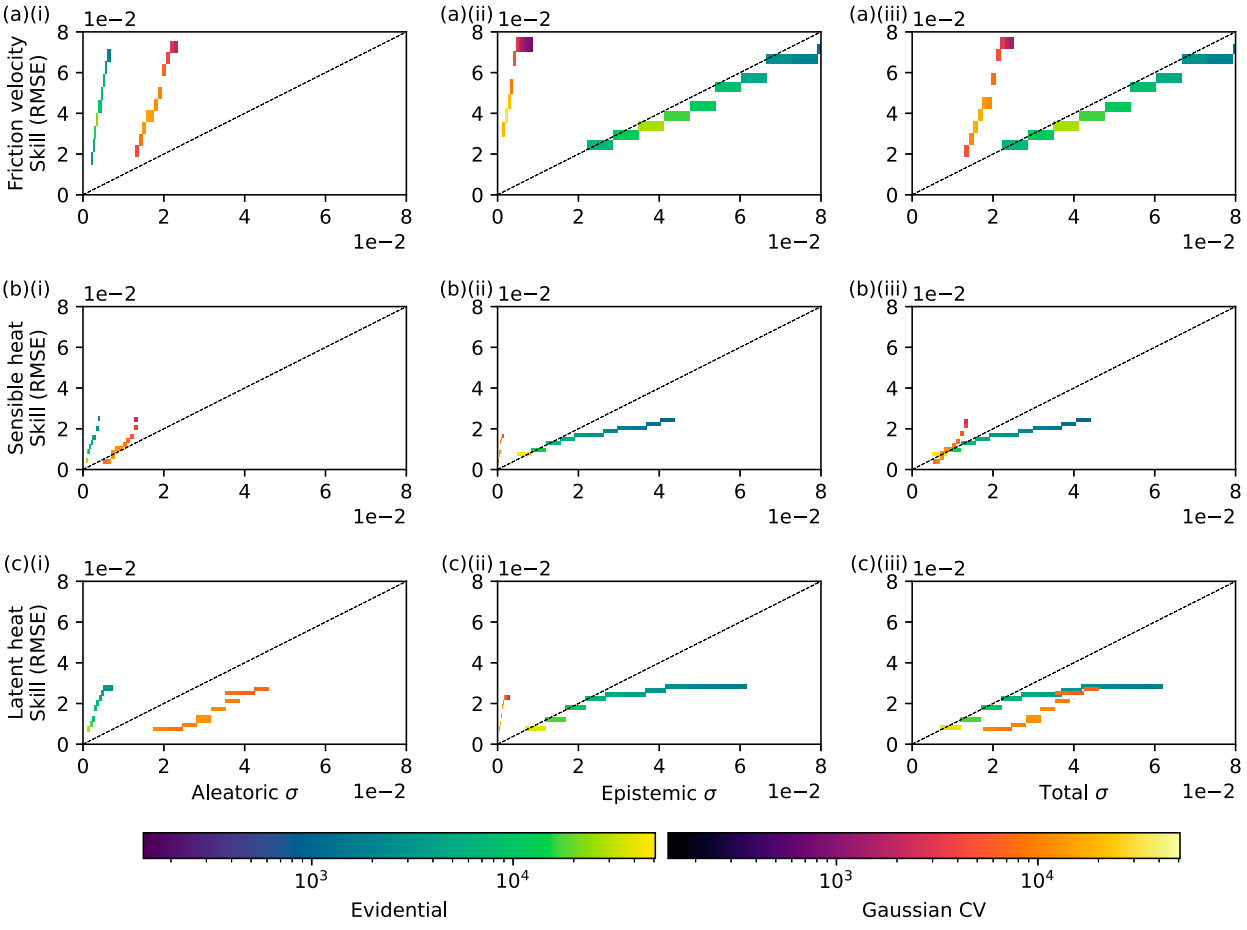

FIG. 9. The spread–skill relationship is depicted using a 2D histogram, illustrating the relationship between the standard deviation σ and RMSE for the three single-task evidential models. Each row in (a)–(c) represents a specific model task, while columns (i)–(iii) display the results for aleatoric, epistemic, and total uncertainties. Each subpanel features a 1-to-1 line (dashed), and all panels share the color bar indicating the count of each 2D bin.

speeds, and more by the wind speed for positive gradients (stable regime). The predicted uncertainties (Figs. 12b–d) follow similar patterns to the sensible heat flux predictions. Epistemic uncertainty increases faster than aleatoric uncertainty in the unstable regime. Aleatoric uncertainty shows higher relative values in the stable regime, which aligns with the 10-m temperature gradient providing less information about the surface sensible heat flux under more stratified and less turbulent conditions.

5. Discussion

For the precipitation-type classification task, the ENN model achieved comparable accuracy to traditional classifiers while simultaneously quantifying the aleatoric uncertainty arising from the inherent biases and data quality issues in the training data (especially inconsistency in crowd-sourced observations) and the epistemic uncertainty stemming from the model's generalization errors aided by gaps in training data. The aleatoric uncertainty generally exceeded the epistemic uncertainty in this problem, underscoring the benefit of decomposing and analyzing the different sources of uncertainty. For example, identifying irreducible aleatoric uncertainty can guide users to data quality control measures or feature engineering techniques to better separate the data. Identification of high epistemic uncertainty can potentially highlight data gaps that can reduce uncertainty with more targeted data collection. The close alignment between regions of high DST uncertainty and peaks in epistemic uncertainty highlights how different uncertainty metrics can capture related aspects of model confidence. Furthermore, the discard test validated the effectiveness of using the uncertainty estimates to filter out unreliable predictions, thereby improving operational reliability.

For categorical problems like the winter precipitation-type classification, the use of DST for quantifying uncertainty offers additional benefits. It allows for representation and reasoning about uncertain or conflicting evidence, which is useful in real-world datasets with ambiguous or uncertain ground truth labels. DST also enables the principled combination of evidence from multiple sources, which can be beneficial when dealing with different data types or integrating information

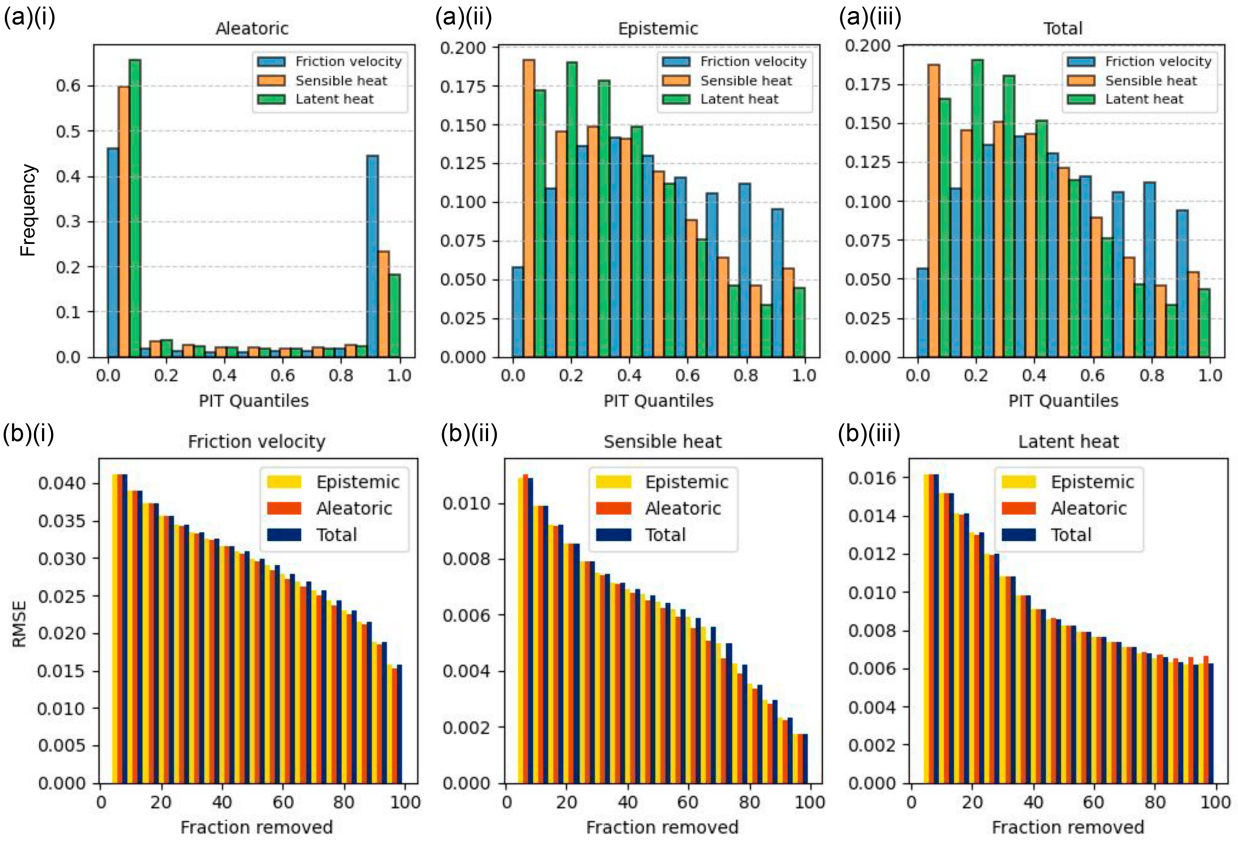

FIG. 10. (a) PIT histograms were generated for the three tasks using (i) aleatoric, (ii) epistemic, and (iii) total uncertainties. (b) The discard-test diagrams illustrate the relationship between the fraction of data points removed from the test set and the remaining subset's RMSE for the three single-task models shown in (i)–(iii).

from diverse sensors or models. The ability of an NN to essentially abstain from making a prediction is valuable in situations with limited information or incomplete evidence, allowing for a more nuanced representation of uncertainty.

For the surface-layer flux regression problem, proper calibration emerged as a crucial factor for obtaining meaningful uncertainty estimates. While the ENN's overall performance was comparable to a deterministic MLP, occasional instances of unrealistic uncertainty values emphasize the need for cautious interpretation and further improvements in calibration. Architectural choices and dropout rates significantly impacted the uncertainty characteristics, suggesting the importance of careful model design and hyperparameter tuning. The time-of-day analysis revealed sensible patterns in the variation of epistemic uncertainty, with higher uncertainty during daytime hours when modeling turbulent transport and radiation processes is more challenging. This observation provides reassurance about the reliability of the uncertainty estimates and their ability to capture fundamental governing processes.

Ensemble and evidential methods for regression tasks show distinct differences in their treatment of uncertainty. Evidential models exhibited higher epistemic uncertainty, while ensemble methods emphasized aleatoric uncertainty. Evidential models exhibited higher epistemic uncertainty, likely resulting from more degrees of freedom in the model architecture than ensemble methods. In contrast, ensemble methods emphasized aleatoric rather than epistemic uncertainty, likely due to the stronger regularization compared with evidential models. Applying the LoTV to the normal-inverse-gamma distribution, we find that in evidential models, epistemic uncertainty increases when the model is more uncertain and adapts readily to new data. Conversely, ensembles tend to show higher aleatoric uncertainty, as individual models capture inherent data noise, and model disagreement (epistemic uncertainty) remains smaller. Notably, the evidential model's total uncertainty was well-calibrated across tasks, whereas the ensemble was only calibrated for sensible heat predictions. This suggests that evidential models offer more flexibility in representing uncertainty but require careful regularization to balance the exploration of hypothesis space. In contrast, ensemble methods, while constrained by fixed data structures and a limited number of models, tend to overemphasize aleatoric uncertainty, potentially at the expense of accurately capturing epistemic uncertainty. Future exploration could consider how to further optimize these methods. For ensembles, introducing regularization coefficients to the loss could potentially widen or adjust the epistemic spread. For evidential models, stronger regularization might limit the exploration of the hypothesis space, reducing the epistemic magnitude.

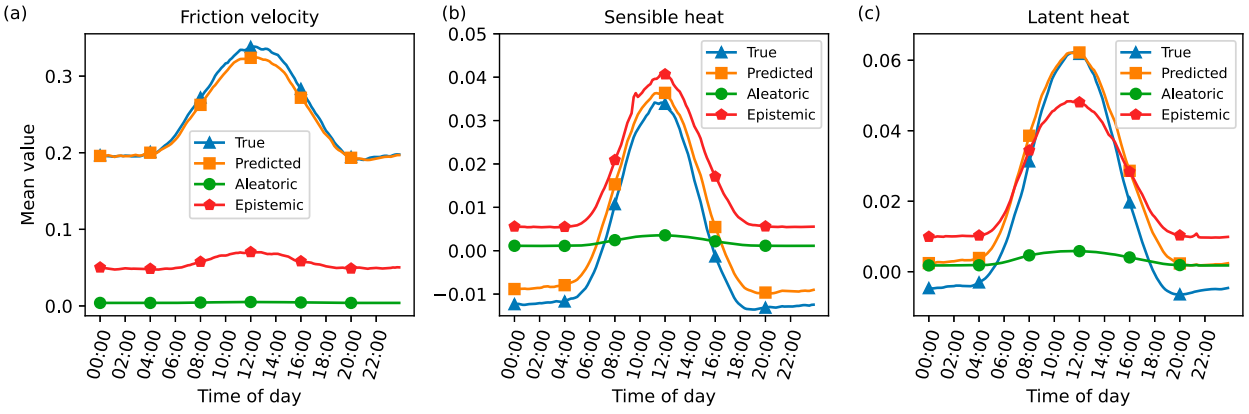

FIG. 11. The mean values of the (a) friction velocity, (b) sensible heat, and (c) latent heat in 10-min increments. The curves displayed in each panel represent the truth, predicted, aleatoric, and epistemic quantities. All results are derived from the test dataset.

These findings highlight the strengths and limitations of both ensemble and evidential approaches and underscore the practical challenges in achieving well-calibrated uncertainty estimates. A key limitation highlighted in this study is the extensive hyperparameter tuning and computational cost required for effective uncertainty calibration (see section SVI in the online supplemental material), particularly for the evidential regression model. This underscores the greater challenges associated with calibrating uncertainties for regression tasks compared to classification problems. Calibrating ML

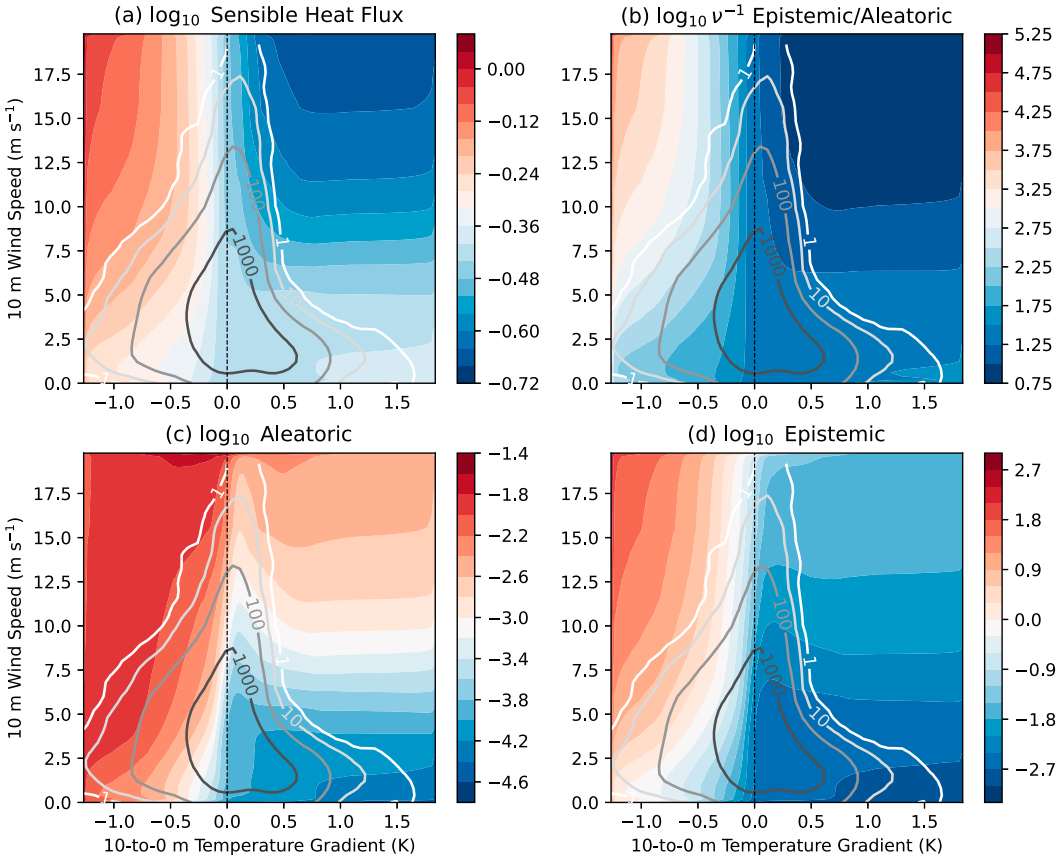

FIG. 12. Each subfigure illustrates the relationship between two of the kinetic sensible heat flux evidential model inputs and the predictions or uncertainties. The gray overlay contours indicate the number of training examples that have a corresponding temperature gradient and wind speed value.

models to account for epistemic uncertainty typically requires either a dedicated calibration/validation dataset or prior assumptions derived from similar data. While this requirement is more apparent for post hoc calibration methods like isotonic regression, it also applies to evidential methods through the tuning of the KL divergence term in the loss function. We have found that the raw target predictions themselves are not very sensitive to the KL divergence term, and thus, we recommend performing hyperparameter tuning with a fixed KL divergence term first and then calibrating uncertainty separately with the tuned model. However, if the trained ENN is applied to a dataset with a distribution shift, such as transitioning from training on analysis data to applying the model to forecast data, the evidential uncertainty estimates may become inherently underdispersive. Further research is needed to develop techniques for adjusting uncertainty estimates based on the nature and extent of the domain shift. While we used RMSE versus σ_{tot} relationships and PIT histograms for regression calibration, equivalent metrics for discrete classification problems are lacking. Future work should develop robust calibration metrics for classification tasks, enabling comprehensive comparison of uncertainty quantification methods across problem types.

6. Conclusions

In conclusion, this study demonstrates the potential of EDL as an effective technique for predictive modeling and UQ in weather and climate applications, for both classification and regression. The approach synergistically integrates the capabilities of probabilistic modeling with the representational power of deep NNs. This enables the models to produce well-calibrated estimates of uncertainty without ensembles or sampling, overcoming the limitations of standard deep learning approaches. The ability to quantify different sources of uncertainty provides valuable insights into model reliability and limitations of the training data.

The decomposition of aleatoric and epistemic uncertainties facilitates detailed analysis to identify challenging prediction cases and opportunities for model improvements. With calibrated uncertainty estimates, ENNs have the potential to enhance understanding of forecast reliability and inform critical decision-making across various meteorological domains, from real-time severe event prediction to long-term climate projections. Given the representational flexibility, computational efficiency, and uncertainty quantification capabilities, EDL shows promise for tackling a diverse array of prediction and uncertainty estimation problems in the atmospheric and climate sciences.

7. Software development

The Machine Integration and Learning for Earth Systems (MILES) group Generalized Uncertainty for Earth System Science (GUESS) package (MILES-GUESS) provides tools for estimating and analyzing different sources of uncertainty in Earth system science applications. Users working in weather and climate can leverage MILES-GUESS to train neural network models that quantify multiple uncertainty types like aleatoric, epistemic, and DST uncertainties.

The package contains layers and losses for EDL, allowing users to build neural networks that output distributions over targets rather than just point estimates. This enables estimating both aleatoric uncertainty from noise in the data and epistemic uncertainty from model limitations. The code also supports MC dropout ensembles for epistemic uncertainty quantification.

Once models are trained with MILES-GUESS, the package provides a range of analysis and visualization tools tailored for Earth system applications. These include PIT calibration analysis, spread-error diagrams, coverage curves, and more. The code is primarily written in Python and is designed to integrate seamlessly with common Earth science workflows based on TensorFlow/Keras and PyTorch. MILES-GUESS is accessible online (https://github.com/ai2es/miles-guess) where example Jupyter Notebooks demonstrate how to use the package.

Acknowledgments. This material is based upon work supported by the National Science Foundation under Grant RISE-2019758 and by the NSF National Center for Atmospheric Research, which is a major facility sponsored by the National Science Foundation under Cooperative Agreement 1852977. We would like to acknowledge high-performance computing support from Cheyenne and Casper (Computational and Information Systems Laboratory 2020) provided by NCAR's Computational and Information Systems Laboratory, sponsored by the National Science Foundation. MJM was supported by the University of Maryland Grand Challenges Program and the U.S. Department of Energy (DOE), Office of Science, RGMA component of the EESM program under Award DE-SC0022070 and National Science Foundation IA 1947282.

Data availability statement. All datasets used in this study are available at https://doi.org/10.5281/zenodo.8368187. The MILES-GUESS package is archived at https://doi.org/10.5281/zenodo.10729801. The surface layer model weights and evaluation data are archived at https://doi.org/10.5281/zenodo.13774771. The winter precipitation-type model weights and evaluation data are archived at https://doi.org/10.5281/zenodo.13776835.

REFERENCES

Abdar, M., and Coauthors, 2021: A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Inf. Fusion*, **76**, 243–297, https://doi.org/10.1016/j.inffus.2021.05.008.

Amini, A., W. Schwarting, A. Soleimany, and D. Rus, 2020: Deep evidential regression. *NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems*, Curran Associates Inc., 14 927–14 937, https://dl.acm.org/doi/abs/10.5555/3495724.3496975.

Angelopoulos, A. N., and S. Bates, 2022: A gentle introduction to conformal prediction and distribution-free uncertainty

quantification. arXiv, 2107.07511v6, https://doi.org/10.48550/arXiv.2107.07511.

Barnes, E. A., and R. J. Barnes, 2021: Controlled abstention neural networks for identifying skillful predictions for regression problems. *J. Adv. Model. Earth Syst.*, **13**, e2021MS002575, https://doi.org/10.1029/2021MS002575.

Bauer, P., A. Thorpe, and G. Brunet, 2015: The quiet revolution of numerical weather prediction. *Nature*, **525**, 47–55, https://doi.org/10.1038/nature14956.

Bosveld, F. C., P. Baas, A. C. M. Beljaars, A. A. M. Holtslag, J. V.-G. de Arellano, and B. J. H. van de Wiel, 2020: Fifty years of atmospheric boundary-layer research at Cabauw serving weather, air quality and climate. *Bound.-Layer Meteor.*, **177**, 583–612, https://doi.org/10.1007/s10546-020-00541-w.

Breiman, L., 1996: Bagging predictors. *Mach. Learn.*, **24**, 123–140, https://doi.org/10.1007/BF00058655.

Bremnes, J. B., 2020: Ensemble postprocessing using quantile function regression based on neural networks and Bernstein polynomials. *Mon. Wea. Rev.*, **148**, 403–414, https://doi.org/10.1175/MWR-D-19-0227.1.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78** (1), 1–3, https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Casella, G., and R. L. Berger, 2002: *Statistical Inference*. 2nd ed. Duxbury Press, 686 pp.

Chapman, W. E., L. Delle Monache, S. Alessandrini, A. C. Subramanian, F. M. Ralph, S.-P. Xie, S. Lerch, and N. Hayatbini, 2022: Probabilistic predictions from deterministic atmospheric river forecasts with deep learning. *Mon. Wea. Rev.*, **150**, 215–234, https://doi.org/10.1175/MWR-D-21-0106.1.

Computational and Information Systems Laboratory, 2020: Cheyenne: HPE/SGI ICE XA System (NCAR Community Computing). NCAR, https://doi.org/10.5065/D6RX99HX.

Delaunay, A., and H. M. Christensen, 2022: Interpretable deep learning for probabilistic MJO prediction. *Geophys. Res. Lett.*, **49**, e2022GL098566, https://doi.org/10.1029/2022GL098566.

Dempster, A. P., 1968: A generalization of Bayesian inference. *J. Roy. Stat. Soc.*, **30B**, 205–232, https://doi.org/10.1111/j.2517-6161.1968.tb00722.x.

Dietterich, T. G., 2000: Ensemble methods in machine learning. *MCS '00: Proceedings of the First International Workshop on Multiple Classifier Systems*, Springer-Verlag, 1–15, https://dl.acm.org/doi/10.5555/648054.743935.

Foster, D., D. J. Gagne II, and D. B. Whitt, 2021: Probabilistic machine learning estimation of ocean mixed layer depth from dense satellite and sparse in situ observations. *J. Adv. Model. Earth Syst.*, **13**, e2021MS002474, https://doi.org/10.1029/2021MS002474.

Gal, Y., and Z. Ghahramani, 2016: Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. *Proc. 33rd Int. Conf. on Machine Learning*, New York, NY, JMLR.org, 1050–1059, https://dl.acm.org/doi/10.5555/3045390.3045502.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, 2013: *Bayesian Data Analysis*. Vol. 2. Chapman and Hall/CRC, 552 pp.

Ghazvinian, M., Y. Zhang, D.-J. Seo, M. He, and N. Fernando, 2021: A novel hybrid artificial neural network–Parametric scheme for postprocessing medium-range precipitation forecasts. *Adv. Water Resour.*, **151**, 103907, https://doi.org/10.1016/j.advwatres.2021.103907.

Gneiting, T., A. E. Raftery, A. H. Westveld III, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.*, **133**, 1098–1118, https://doi.org/10.1175/MWR2904.1.

——, F. Balabdaoui, and A. E. Raftery, 2007: Probabilistic forecasts, calibration and sharpness. *J. Roy. Stat. Soc.*, **69B**, 243–268, https://doi.org/10.1111/j.1467-9868.2007.00587.x.

Gordon, E. M., and E. A. Barnes, 2022: Incorporating uncertainty into a regression neural network enables identification of decadal state-dependent predictability in CESM2. *Geophys. Res. Lett.*, **49**, e2022GL098635, https://doi.org/10.1029/2022GL098635.

Guillaumin, A. P., and L. Zanna, 2021: Stochastic-deep learning parameterization of ocean momentum forcing. *J. Adv. Model. Earth Syst.*, **13**, e2021MS002534, https://doi.org/10.1029/2021MS002534.

Hamill, T. M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.*, **129**, 550–560, https://doi.org/10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2.

Haupt, S. E., W. Chapman, S. V. Adams, C. Kirkwood, J. S. Hosking, N. H. Robinson, S. Lerch, and A. C. Subramanian, 2021: Towards implementing artificial intelligence post-processing in weather and climate: Proposed actions from the Oxford 2019 workshop. *Philos. Trans. Roy. Soc.*, **A379**, 20200091, https://doi.org/10.1098/rsta.2020.0091.

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

Herman, G. R., and R. S. Schumacher, 2018: Money doesn't grow on trees, but forecasts do: Forecasting extreme precipitation with random forests. *Mon. Wea. Rev.*, **146**, 1571–1600, https://doi.org/10.1175/MWR-D-17-0250.1.

Hoffman, M. D., D. M. Blei, C. Wang, and J. Paisley, 2013: Stochastic variational inference. *J. Mach. Learn. Res.*, **14**, 1303–1347.

Jøsang, A., 2018: *Subjective Logic: A Formalism for Reasoning Under Uncertainty*. Springer, 337 pp.

Karpatne, A., and Coauthors, 2017: Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Trans. Knowl. Data Eng.*, **29**, 2318–2331, https://doi.org/10.1109/TKDE.2017.2720168.

Kendall, A., and Y. Gal, 2017: What uncertainties do we need in Bayesian deep learning for computer vision? *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, Curran Associates Inc., 5580–5590, https://dl.acm.org/doi/10.5555/3295222.3295309.

Koenker, R., 2005: *Quantile Regression*. Cambridge University Press, 368 pp.

Kotz, S., N. Balakrishnan, and N. L. Johnson, 2004: *Continuous Multivariate Distributions.* Vol. 1, *Models and Applications*. John Wiley and Sons, 753 pp.

Lakshminarayanan, B., A. Pritzel, and C. Blundell, 2017: Simple and scalable predictive uncertainty estimation using deep ensembles. *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, Curran Associates Inc., 6405–6416, https://dl.acm.org/doi/10.5555/3295222.3295387.

Leith, C. E., 1974: Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.*, **102**, 409–418, https://doi.org/10.1175/1520-0493(1974)102<0409:TSOMCF>2.0.CO;2.

Liu, J. Z., Z. Lin, S. Padhy, D. Tran, T. Bedrax-Weiss, and B. Lakshminarayanan, 2020: Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *NIPS'20: Proceedings of the 34th International*

Conference on Neural Information Processing Systems, Curran Associates Inc., 7498–7512, https://dl.acm.org/doi/10.5555/3495724.3496353.

MacKay, D. J., 2003: *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 628 pp.

McCandless, T., D. J. Gagne, B. Kosović, S. E. Haupt, B. Yang, C. Becker, and J. Schreck, 2022: Machine learning for improving surface-layer-flux estimates. *Bound.-Layer Meteor.*, **185**, 199–228, https://doi.org/10.1007/s10546-022-00727-4.

McGovern, A., K. L. Elmore, D. J. Gagne II, S. E. Haupt, C. D. Karstens, R. Lagerquist, T. Smith, and J. K. Williams, 2017: Using artificial intelligence to improve real-time decision-making for high-impact weather. *Bull. Amer. Meteor. Soc.*, **98**, 2073–2090, https://doi.org/10.1175/BAMS-D-16-0123.1.

Meinert, N., and A. Lavin, 2021: Multivariate deep evidential regression. arXiv, 2104.06135v4, https://doi.org/10.48550/arXiv.2104.06135.

——, J. Gawlikowski, and A. Lavin, 2023: The unreasonable effectiveness of deep evidential regression. *Proc. AAAI Conf. Artif.* Intell., **37**, 9134–9142, https://doi.org/10.1609/aaai.v37i8.26096.

Muñoz-Esparza, D., C. Becker, J. A. Sauer, D. J. Gagne II, J. Schreck, and B. Kosović, 2022: On the application of an observations-based machine learning parameterization of surface layer fluxes within an atmospheric large-eddy simulation model. *J. Geophys. Res. Atmos.*, **127**, e2021JD036214, https://doi.org/10.1029/2021JD036214.

Murphy, A. H., 1973: A new vector partition of the probability score. *J. Appl. Meteor.*, **12**, 595–600, https://doi.org/10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2.

Murphy, K. P., 2007: *Conjugate Bayesian Analysis of the Gaussian Distribution*. University of British Columbia, 29 pp.

Nadav-Greenberg, L., and S. L. Joslyn, 2009: Uncertainty forecasts improve decision making among nonexperts. *J. Cognit. Eng. Decis. Making*, **3**, 209–227, https://doi.org/10.1518/155534309X474460.

Neal, R. M., 2012: *Bayesian Learning for Neural Networks*. Lecture Notes in Statistics, Vol. 118. Springer, 204 pp.

Nix, D. A., and A. S. Weigend, 1994: Estimating the mean and variance of the target probability distribution. *Proc. 1994 IEEE Int. Conf. on Neural Networks (ICNN'94)*, Orlando, FL, Institute of Electrical and Electronics Engineers, 55–60, https://ieeexplore.ieee.org/document/374138.

Oh, D., and B. Shin, 2022: Improving evidential deep learning via multi-task learning. *Proc. AAAI Conf. Artif.* Intell., **36**, 7895–7903, https://doi.org/10.1609/aaai.v36i7.20759.

Ovadia, Y., and Coauthors, 2019: Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Curran Associates Inc., 14 003–14 014, https://dl.acm.org/doi/abs/10.5555/3454287.3455541.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.*, **133**, 1155–1174, https://doi.org/10.1175/MWR2906.1.

Rasmussen, C. E., and C. K. I. Williams, 2006: *Gaussian Processes for Machine Learning*. The MIT Press, 248 pp.

Rasp, S., and S. Lerch, 2018: Neural networks for postprocessing ensemble weather forecasts. *Mon. Wea. Rev.*, **146**, 3885–3900, https://doi.org/10.1175/MWR-D-18-0187.1.

Romano, Y., E. Patterson, and E. J. Candes, 2019: Conformalized quantile regression. *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Curran Associates Inc., 3543–3553, https://dl.acm.org/doi/10.5555/3454287.3454605.

Scheuerer, M., M. B. Switanek, R. P. Worsnop, and T. M. Hamill, 2020: Using artificial neural networks for generating probabilistic subseasonal precipitation forecasts over California. *Mon. Wea. Rev.*, **148**, 3489–3506, https://doi.org/10.1175/MWR-D-20-0096.1.

Schulz, B., and S. Lerch, 2022: Machine learning methods for postprocessing ensemble forecasts of wind gusts: A systematic comparison. *Mon. Wea. Rev.*, **150**, 235–257, https://doi.org/10.1175/MWR-D-21-0150.1.

Sensoy, M., L. Kaplan, and M. Kandemir, 2018: Evidential deep learning to quantify classification uncertainty. *NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems*, Curran Associates Inc., 3183–3193, https://dl.acm.org/doi/10.5555/3327144.3327239.

——, V. Deli'c, and L. Kaplan, 2020: A review of uncertainty quantification in deep learning: Techniques, applications and challenges. arXiv, 2011.06225v4, https://doi.org/10.48550/arXiv.2011.06225.

Shaker, M. H., and E. Hüllermeier, 2020: Aleatoric and epistemic uncertainty with random forests. *Advances in Intelligent Data Analysis XVIII: 18th International Symposium on Intelligent Data Analysis, IDA 2020, Konstanz, Germany, April 27–29, 2020, Proceedings*, Springer-Verlag, 444–456, https://dl.acm.org/doi/10.1007/978-3-030-44584-3_35.

Shepherd, J. G., 2009: Geoengineering the climate: Science, governance and uncertainty. 98 pp., https://royalsociety.org/-/media/policy/publications/2009/8693.pdf.

Soleimany, A. P., A. Amini, S. Goldman, D. Rus, S. N. Bhatia, and C. W. Coley, 2021: Evidential deep learning for guided molecular property prediction and discovery. *ACS Cent. Sci.*, **7**, 1356–1367, https://doi.org/10.1021/acscentsci.1c00546.

Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, 2014: Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, **15**, 1929–1958.

Stankeviciute, K., A. M. Alaa, and M. van der Schaar, 2021: Conformal time-series forecasting. *Advances in Neural Information Processing Systems 34*, M. Ranzato et al., Eds., Curran Associates, Inc., 6216–6228, https://proceedings.neurips.cc/paper/2021/hash/312f1ba2a72318edaaa995a67835fad5-Abstract.html.

Ulmer, D., C. Hardmeier, and J. Frellsen, 2021: Prior and posterior networks: A Survey on evidential deep learning methods for uncertainty estimation. *Trans. Mach. Learn. Res.*, **2023**, 1–48.

Vannitsem, S., D. S. Wilks, and J. Messner, 2018: *Statistical Postprocessing of Ensemble Forecasts*. 1st ed. Elsevier, 362 pp.

——, and Coauthors, 2021: Statistical postprocessing for weather forecasts: Review, challenges, and avenues in a big data world. *Bull. Amer. Meteor. Soc.*, **102**, E681–E699, https://doi.org/10.1175/BAMS-D-19-0308.1.