Uncertainty Quantification in Deep Learning: Bayesian Mixture Density Networks for Carbon Flux Prediction

by

Anish Dulal

A thesis accepted and approved in partial fulfillment of the

requirements for the degree of

Master of Science

in Computer Science

Thesis Committee:

Jake Searcy, Chair

Daniel Lowd, Core Member

Thien Huu Nguyen, Core Member

University of Oregon

Fall 2025

THESIS ABSTRACT

Anish Dulal

Master of Science in Computer Science

Title: Uncertainty Quantification in Deep Learning: Bayesian Mixture Density Networks for Carbon Flux Prediction

Machine-learning approaches are central to tower-to-regional upscaling of terrestrial carbon fluxes, yet their outputs are often used as if a single predictive variance were sufficient for risk management and crediting decisions. This thesis argues that for high-stakes applications, we must know not only how uncertain predictions are but also why. I formalize a three-way decomposition of predictive uncertainty into aleatoric, compositional, and epistemic components (ACE) for regression problems, and derive estimators of ACE from a Bayesian mixture density network (BMDN). The network couples a deterministic feature encoder with a Bayesian last layer and a Gaussian mixture output, so that within-component spread, between-component separation, and posterior dispersion of the mean maps directly to A, C, and E. Finally, I propose an ACE-based decision recipe that helps determine when scalable carbon flux monitoring is reliable, and provides suggestions on whether to gather data or add new variables to improve carbon flux predictions.

A synthetic multimodal regression benchmark with known ground-truth ACE is used to stress-test the decomposition and to examine how mixture complexity, prior scale, and training-set design influence calibration and proper scoring rules. I then apply the framework to net ecosystem exchange from AmeriFlux towers conditioned on Landsat 8/9 reflectances, meteorological, and ancillary land-surface descriptors. The BMDN attains accuracy comparable to strong baselines while improving probabilistic calibration under both temporal extrapolation and spatial transfer to unseen sites, and the ACE fields reveal systematic patterns across biomes and feature ablations, including signatures of noisy flux regimes, missing covariates, and data-sparse regions.

This thesis includes previously published co-authored material.

ACKNOWLEDGEMENTS

I am grateful to my advisor, Professor Jake Searcy, for his guidance, patience, and encouragement throughout this work. I also thank Professors Daniel Lowd and Thien Huu Nguyen for serving on my committee and for their feedback. This thesis benefited from discussions with collaborators and colleagues in the Department of Ecology and from the broader AmeriFlux community, who maintain the tower network and data products on which this work relies. I am thankful to my labmates and friends for creating a supportive environment in which to learn, experiment, and write. This material is based upon work supported by the National Science Foundation under Grant No. 2319597.

TABLE OF CONTENTS

Chapter Page

LIST OF FIGURES

LIST OF TABLES

Table Page

CHAPTER I
INTRODUCTION

1.1 Motivation and problem context

Climate change remains one of the most pressing challenges of the 21st century. Recent assessments from the Global Carbon Project indicate that the remaining carbon budget for a 50% chance of limiting global warming to 1.5 °C has fallen to roughly 65 Gt C ($\approx$ 235 Gt CO_2), which is equivalent to about six years of emissions at 2024 levels Friedlingstein et al. (2023); Smith et al. (2024). Atmospheric CO_2 concentrations are projected to reach $\sim$422.45 ppm in 2024, about 52% higher than pre-industrial levels, and both ocean and land carbon sinks show signs of stress Friedlingstein et al. (2023); IPCC (2023). While decarbonization efforts focused on reducing fossil fuel use are essential, they are unlikely to be sufficient on their own to achieve this ambitious target. Scaling up natural climate solutions (NCS), which involve the conservation, restoration, and improved management of ecosystems to enhance carbon storage and avoid emissions, is therefore widely viewed as an essential complement to rapid fossil fuel phaseout Chausson et al. (2020); Griscom et al. (2017); Seddon et al. (2020).

Accurate and transparent quantification of terrestrial carbon fluxes is critical both for monitoring the performance of natural climate solutions and for ensuring integrity in carbon markets. Flux estimates underpin our understanding of ecosystem carbon dynamics and inform climate models, national greenhouse gas inventories, and crediting schemes IPCC (2023); Smith et al. (2024). Many voluntary standards and regulatory programs explicitly couple credit issuance to quantified uncertainty: projects often face uncertainty thresholds that trigger an automatic discount in credited tons or, in some cases, prevent listing if relative uncertainty is too high Haya et al. (2020); Teo et al. (2023). Since credited tons depend directly on net fluxes and their uncertainty, it is not enough to estimate average fluxes. Practitioners need to know how uncertain those estimates are, and how that uncertainty is structured across space, time, and drivers.

Yet quantifying carbon fluxes at scale remains challenging. Eddy covariance (EC) towers measure turbulent fluxes of heat, water vapor, and carbon dioxide at high temporal resolution, but their spatial footprints are dynamic and depend on atmospheric stability, wind direction, and surface heterogeneity Baldocchi (2003); Chu et al. (2021); Kljun, Calanca, Rotach, and Schmid (2015); Schmid (2002); Vesala

et al. (2008). Installing and maintaining towers is expensive, and the global network remains sparse and unevenly distributed across biomes and climate zones. This limits our ability to monitor carbon fluxes across large, heterogeneous landscapes and to evaluate how representative tower sites are of their surroundings. Section 2.2 in Chapter II reviews EC instrumentation, footprints, and global tower networks in more detail.

Remote sensing addresses part of this challenge by providing spatially continuous information on surface state and vegetation structure. The Landsat program, and in particular Landsat 8, provides global moderate resolution multispectral observations that have become a workhorse for ecological and forestry applications Morfitt et al. (2015); NASA Landsat Science Team (n.d.); Roy et al. (2014); Vermote, Justice, Claverie, and Franch (2016); Wulder et al. (2019). These data have been widely combined with flux tower measurements and meteorological drivers in machine learning models to upscale local flux observations to regional and global grids Jung et al. (2019, 2020); Tramontana et al. (2016); Zeng et al. (2020). Throughout this thesis we follow a similar strategy: we combine Landsat imagery with meteorological and ancillary variables which can be used to extrapolate point-wise flux measurements to larger areas. Technical details of the remote sensing and meteorological drivers used here are reviewed in Section 2.3.

1.2 From point estimates to predictive distributions

Machine learning models have been used extensively to relate remote sensing data and meteorological variables to carbon fluxes. Early approaches such as Random Forests and feedforward neural networks learn nonlinear relationships between predictors and fluxes but typically provide only point estimates and do not quantify predictive uncertainty. A standard neural network, for example, maps inputs to a single deterministic output and does not express how confident the model is in its prediction Lakshminarayanan, Pritzel, and Blundell (2017). In carbon markets and climate risk analysis, however, decision makers often need explicit uncertainty estimates. Many crediting programs and buyers discount or withhold credits when estimated uncertainty is large, in order to manage the risk of over-crediting and non-compliance Haya et al. (2020); Teo et al. (2023). Uncertainty estimation is therefore not an optional add-on. It is a core requirement: users need to understand not only what the model predicts, but how reliable those predictions are and why.

Two broad types of uncertainty are usually distinguished in machine learning and related fields Depeweg, Hernández-Lobato, Doshi-Velez, and Udluft (2018); Kendall and Gal (2017); Walker et al. (2003):

Epistemic uncertainty reflects limited knowledge in the model. It captures uncertainty about parameters and, more broadly, limitations of the model class and feature set. In principle it can be reduced by collecting more or more informative data, improving the model structure, or including additional relevant drivers.

Aleatoric uncertainty reflects variability that is irreducible at the scale of the model. It includes measurement noise, unresolved sub-grid variability, and stochastic processes that remain even if the model and data were perfect.

In this thesis we argue that these two categories are not always sufficient to describe the uncertainty structure that arises in multivalued regression problems. We introduce a third component, **compositional uncertainty**, which captures ambiguity between multiple plausible patterns in the target when the same observed drivers can lead to distinct flux levels. Intuitively, compositional uncertainty arises when different dynamical regimes or land-surface states share similar spectral and meteorological signatures, so that single-valued predictors are forced to average over multiple plausible outcomes.

Chapter III formalizes this three-way ACE (aleatoric, compositional, epistemic) decomposition and derives it from the law of total variance. In this introduction we focus on the high-level motivation and on how ACE connects to concrete predictive models.

A simple probabilistic extension of point-estimate regression is Gaussian regression. Instead of assuming that the model produces a single "best guess" value, we assume that, for a given input, the target flux is drawn from a normal distribution characterized by a mean and a variance. The mean represents the expected flux for that set of drivers and the variance represents the spread around that mean. In the homoscedastic case this variance is treated as constant across inputs, whereas in the heteroscedastic case the variance is allowed to depend on the input so that the model can express that some situations are intrinsically noisier than others. Heteroscedastic Gaussian regression networks implement this idea by predicting both a mean and an input-dependent variance, and in doing so they provide a natural representation of aleatoric uncertainty Kendall and Gal (2017). However, these models still impose a

unimodal, approximately Gaussian response. They can capture within-regime noise, but multi-regime structure is collapsed into a single mode, motivating a separate component to track multivalued structure.

1.3 Mixture models and mixture density networks

A natural way to model multimodal conditional distributions is through mixture models. In a Gaussian mixture model, the target is assumed to arise from a finite mixture of Gaussian components, each with its own mean and variance, and each weighted by a mixture probability. For a given input, each component can be interpreted as a different plausible regime or pattern, while the mixture weights represent how likely each regime is.

Mixture density networks (MDNs) provide a probabilistic framework for modeling such complex, multimodal conditional distributions. An MDN augments a neural network with a mixture model. Instead of outputting a single value, the network predicts the parameters (means, variances, and mixture weights) of a Gaussian mixture conditioned on the input Bishop (1994, 2006). Each mixture component reflects one plausible output, while the mixing weights represent the model's belief about how likely each component is given the input. Through this probabilistic output, MDNs can capture both within-regime variability and multi-regime structure, making them well suited to multivalued problems.

This mixture structure connects directly to the ACE decomposition. The mixture-weighted within-component variance naturally corresponds to the aleatoric component, capturing noise and sub-grid variability within each regime. The variance of the component means around the mixture mean corresponds to the compositional component, capturing ambiguity between multiple plausible regimes. The remaining epistemic component describes uncertainty about the mixture parameters themselves when data are limited or the model is misspecified.

MDNs and related mixture-based regressors have been applied in a range of domains where predictive distributions and uncertainty matter, including control, photonics, medical applications, hydrology, and energy forecasting Hepp, Zierk, and Seitz (2022); Herzallah and Lowe (2004); Men, Yee, Lien, Wen, and Chen (2016); Saranathan et al. (2024); Unni, Yao, and Zheng (2020). These studies demonstrate that mixture-based neural networks can produce calibrated predictive distributions and competitive uncertainty estimates, especially when the target distribution is non-

Gaussian or multimodal. This evidence motivates the use of MDNs for carbon flux estimation, where multi-regime behavior and non-Gaussian tails are common.

Chapter II reviews MDNs and related probabilistic regressors in more detail, and Chapter IV provides the precise parameterization used in this thesis.

1.4 Beyond two-way uncertainty: Bayesian neural networks

Standard MDNs treat network weights as deterministic. They can represent aleatoric and compositional structure in the predictive distribution, but they do not express epistemic uncertainty about the parameters. Bayesian and approximately Bayesian neural network methods address this by placing probability distributions over network weights and using approximate inference schemes such as variational Bayes, Monte Carlo dropout, and deep ensembles to propagate parameter uncertainty to predictions Blundell, Cornebise, Kavukcuoglu, and Wierstra (2015); Gal and Ghahramani (2016); Lakshminarayanan et al. (2017); Neal (1996). In practice, these methods provide different trade-offs between computational cost, flexibility, and the fidelity of the uncertainty estimate.

In this thesis we adopt a three-component decomposition of predictive uncertainty tailored to mixture models. We interpret aleatoric uncertainty as the mixture-weighted within-component variability. We introduce compositional uncertainty to represent multimodality, defined as the between-component variance of the component means. We reserve epistemic uncertainty for uncertainty about the parameters, including weights in the mixture density network and any hyperparameters, which arises from limited or unrepresentative data and from model misspecification.

Chapter III provides formal definitions of the ACE components and shows how they arise from the law of total variance. Chapter IV introduces the Bayesian mixture density network (BMDN) architecture used in this thesis and describes how approximate Bayesian inference in a last-layer treatment is used to obtain practical estimators for the ACE components. The remainder of this introduction emphasises the problem setting, contributions, research questions, and decision context.

1.5 Thesis contributions

This thesis builds upon the above insights to develop a three-way uncertainty-aware framework for estimating carbon fluxes from multispectral Landsat imagery and ancillary environmental data. The key contributions are:

1. **Problem formulation and uncertainty decomposition.** We formulate carbon flux estimation from satellite and meteorological data as a multivalued regression problem, in the sense that the same drivers can correspond to multiple plausible flux values. We introduce the ACE (aleatoric, compositional, epistemic) decomposition of predictive uncertainty, formalised in Chapter III, and relate each component to distinct aspects of the conditional distribution.
2. **Bayesian mixture density network architecture.** We propose a Bayesian mixture density network that combines a deterministic feature extractor, a Bayesian last layer, and a Gaussian mixture output head. Chapter IV details this architecture and the approximate inference scheme used to obtain practical estimators of the ACE components.
3. **Integration and analysis of environmental drivers.** We integrate Landsat 8 spectral bands with meteorological and ancillary variables to capture key drivers of carbon flux dynamics beyond what is visible from reflectance alone. We evaluate how these drivers affect both predictive performance and the ACE decomposition, including ablation-style comparisons to models that use spectral information alone.
4. **Synthetic and real-data validation.** We validate the framework on a controlled synthetic problem with oracle uncertainty, and on AmeriFlux—Landsat data that span multiple ecosystems and climate regimes. In each case we compare the BMDN to simpler probabilistic baselines, including a single-Gaussian ($K = 1$) heteroscedastic regressor and a deterministic MDN. Evaluation uses a combination of point-wise metrics (for example RMSE and R^2), proper scoring rules such as the negative log-likelihood and the continuous ranked probability score, and calibration diagnostics.
5. **Decision-oriented uncertainty analysis.** We show how the ACE components can inform decisions about measurement design, feature collection, and model deployment. In particular, we explore how spatial patterns in A, C, and E can highlight where new towers or additional drivers would most reduce uncertainty, and discuss how compositional and epistemic uncertainty can be translated into discounting or risk management strategies in the context of natural climate solutions and carbon markets.

1.6 Research questions

This work seeks to answer the following questions:

1. **Accuracy and scalability.** Can a mixture density network trained on multispectral Landsat data and environmental variables produce accurate, spatially scalable estimates of carbon fluxes across diverse ecosystems?
2. **Effect of ACE and comparison to baselines.** Does explicitly modeling three types of uncertainty (aleatoric, compositional, and epistemic) improve the quality and interpretability of flux estimates, and how does the resulting BMDN compare to simpler probabilistic baselines such as a single-Gaussian ($K = 1$) heteroscedastic regressor and a deterministic MDN in terms of predictive performance, calibration, and decomposed uncertainty?
3. **Role of environmental drivers.** To what extent does incorporating ancillary environmental drivers, beyond spectral information alone, improve predictive performance and the structure of the ACE decomposition?
4. **Decision support.** How can the ACE components be used to support practical decisions about measurement design, feature collection, and deployment or discounting strategies in natural climate solutions and carbon markets?

1.7 Thesis organization

The remainder of this thesis is organized as follows. Chapter II reviews background on carbon flux estimation, flux tower footprints, remote sensing, and probabilistic regression. Chapter III develops the theory of the three-component ACE (aleatoric, compositional, epistemic) uncertainty decomposition and relates it to the law of total variance. Chapter IV introduces the Bayesian mixture density network architecture and the operational estimators that map model outputs to the ACE components. Chapter V validates the BMDN and the ACE decomposition on a controlled synthetic problem and addresses Research Questions 1–3 in an idealised setting. Chapter VI is adapted from a paper previously published as Dulal and Searcy (2025) in the ICLR 2025 Workshop on Tackling Climate Change with Machine Learning Dulal and Searcy (2025). Chapter VI applies the model to AmeriFlux data and analyzes predictive performance, calibration, and uncertainty patterns across evaluation splits and land-cover types, further addressing Research Questions 1–3 in a real-data setting. Chapter VII concludes the thesis and outlines directions for future research.

CHAPTER II
BACKGROUND AND RELATED STUDY

2.1 Carbon Cycle and Natural Climate Solutions

Understanding terrestrial carbon fluxes is essential for constraining the global carbon budget and informing climate policies Beer et al. (2010); Friedlingstein et al. (2023). At the ecosystem scale, the main biological carbon fluxes are gross primary production (GPP), ecosystem respiration, and their net balance. Gross primary production is the total amount of carbon dioxide (CO_2) taken up by vegetation through photosynthesis per unit area and time; it represents the largest land–atmosphere CO_2 flux in the global carbon cycle and is a primary driver of ecosystem productivity Beer et al. (2010); Zhao, Heinsch, Nemani, and Running (2005). Ecosystem respiration (often denoted R_{eco}) is the sum of autotrophic respiration by plants and heterotrophic respiration by microbes and other organisms that decompose organic matter, and it returns CO_2 from biomass and soils back to the atmosphere Baldocchi (2003).

Net ecosystem exchange (NEE) is defined as the net CO_2 flux between the land surface and the atmosphere over an ecosystem footprint. Following the micrometeorological sign convention used in eddy covariance studies, NEE is typically written

$$\text{NEE} = R_{\text{eco}} - \text{GPP}, \tag{2.1}$$

so that negative values of NEE indicate a net carbon sink (photosynthetic uptake exceeds ecosystem respiration) and positive values indicate a net carbon source to the atmosphere Baldocchi (2003). Throughout this thesis we adopt this sign convention when interpreting tower-based fluxes and model predictions.

Because natural climate solutions aim to enhance GPP, reduce ecosystem respiration, or both, NEE provides a natural diagnostic of whether an ecosystem is acting as a persistent carbon sink or source under a given management or climate regime Chausson et al. (2020); Griscom et al. (2017); Seddon et al. (2020).

2.2 Flux Measurements Using Eddy Covariance

2.2.1 Principle and Instrumentation. The eddy covariance (EC) technique provides direct, continuous measurements of surface-atmosphere exchanges of CO_2, water vapor and energy by correlating high-frequency fluctuations in vertical wind speed with scalar concentrations. An EC tower typically consists of a three-dimensional sonic anemometer to measure wind vectors and an infrared gas

analyzer to measure CO_2 and H_2O concentrations at 10–20 Hz. Fluxes are computed as $F = \overline{w'\,c'}$, where w' and c' denote deviations of the vertical wind component and scalar concentration from their means Baldocchi (2003). The technique integrates fluxes over an upwind source area, known as the footprint, whose shape and extent depend on measurement height, surface roughness, canopy structure, atmospheric stability and wind direction.

2.2.2 Dynamic Footprint and Representativeness. The flux footprint is highly dynamic. For near-uniform terrain, the effective fetch should be roughly ten times the mean canopy height (10H) to satisfy footprint requirements, but in practice heterogeneity in land cover, topography and wind patterns causes the footprint to shift continuously; peak contributions can occur several hundred meters upwind and vary dramatically over diurnal and seasonal cycles Chu et al. (2021); Kljun et al. (2015); Kormann and Meixner (2001); Schmid (2002); Vesala et al. (2008). This complicates the comparison of tower measurements with remote-sensing pixels or land-surface model grid cells. Moreover, towers are typically located in accessible or homogeneous ecosystems, leading to spatial biases in observational coverage. These limitations motivate the development of up-scaling methods that combine tower data with ancillary observations.

2.2.3 Global Tower Networks. Regional networks, such as AmeriFlux, EuroFlux, AsiaFlux and OzFlux, deploy EC towers across different continents. FLUXNET harmonises these data into global releases (e.g., FLUXNET2015), applying consistent preprocessing, gap-filling and quality control Chu et al. (2021); Papale (2020); Pastorello et al. (2020). Although more than 900 sites span diverse climates and vegetation types, coverage remains sparse in tropical forests, mountainous regions and high latitudes. Footprint mismatches and uneven spatial distribution limit direct integration with gridded models and highlight the need for reliable up-scaling techniques.

2.3 Remote Sensing and Ancillary Drivers

2.3.1 Landsat Imagery. Moderate-resolution multispectral observations from the Landsat program provide invaluable information on vegetation structure and function. The Operational Land Imager (OLI) onboard Landsat 8 acquires eleven spectral bands: coastal aerosol, blue, green, red, near-infrared, two shortwave infrared bands, cirrus and two thermal infrared bands. The visible, near-infrared and shortwave infrared bands have a spatial resolution of 30 m, the panchromatic

band has 15 m resolution, and thermal infrared bands have an effective resolution of 100 m Morfitt et al. (2015); Roy et al. (2014). Landsat 8 follows a sun-synchronous orbit at about 705 km altitude, imaging each location every 16 days and capturing approximately 725 scenes per day Roy et al. (2014); (USGS), Aeronautics, and (NASA) (2024); Wulder et al. (2019). OLI uses a push-broom design with 12-bit quantisation, enabling high radiometric precision Morfitt et al. (2015); Vermote et al. (2016). These characteristics make Landsat ideal for deriving vegetation indices that measure vegetation greenness (e.g., NDVI, EVI), monitoring land-cover change and informing flux up-scaling models. Most up-scaling models use this derived NDVI and EVI instead of full spectra.

2.3.2 MODIS and Other Satellite Sensors. The Moderate Resolution Imaging Spectroradiometer (MODIS), carried on NASA's Terra and Aqua satellites, collects data in 36 spectral channels spanning wavelengths from 0.4 to 14.4 µm, providing global coverage every one to two days Justice et al. (1998). This broad spectral range enables studies of vegetation health, land-cover change, ocean biology, sea surface temperature and fire detection Justice et al. (1998). MODIS land products, available at 250–1000 m spatial resolution, include surface reflectance, vegetation indices, land surface temperature and albedo Justice et al. (2002); Schaaf et al. (2002); Zhao et al. (2005). MODIS's high temporal frequency complements the finer spatial resolution but lower revisit rate of Landsat, allowing the derivation of climatologies and the filling of cloud-contaminated gaps. Additional sensors such as Sentinel-2 (10–20 m) and geostationary satellites (e.g., GOES, Himawari) contribute further spatial and temporal detail, though they are not explicitly considered here.

2.3.3 Meteorological and Ancillary Variables. Remote sensing alone cannot capture all factors controlling carbon fluxes because meteorological conditions strongly influence photosynthesis, respiration and canopy conductance. Important drivers include air temperature, photosynthetically active radiation, shortwave and longwave radiation, relative humidity, vapor pressure deficit, wind speed, precipitation and soil moisture. We measure these drivers every 30 minutes from the tower. These variables are typically obtained from ground measurements, atmospheric reanalyses (e.g., ERA5, MERRA-2) or gridded forcing data sets Bosilovich, Robertson, and Takacs (2015); Hersbach et al. (2020). Land-cover classification (e.g., IGBP classes), soil texture and topography provide additional contextual information. Combining remote sensing with meteorological and ancillary

drivers yields feature vectors that capture spatial and temporal patterns influencing carbon fluxes.

2.4 Upscaling Carbon Fluxes: Data-Driven Methods

2.4.1 Early Machine-Learning Approaches. Data-driven up-scaling aims to predict fluxes at unsampled locations by learning statistical relationships between tower observations and environmental covariates. Early studies employed regression trees, random forests, support vector machines and multilayer perceptrons to estimate NEE, GPP and evapotranspiration Breiman (2001). Tree-based ensemble methods and neural networks can model nonlinear interactions and are robust to multicollinearity, yet they produce deterministic point estimates that lack uncertainty quantification Kendall and Gal (2017); Lakshminarayanan et al. (2017). Neural networks trained with mean squared error approximate only the conditional mean of fluxes and cannot capture multi-valued relationships. Quantile regression forests and deep ensembles provide empirical prediction intervals by estimating conditional quantiles or averaging across independent models, but these intervals often lack calibration and do not explicitly disentangle different sources of uncertainty Kendall and Gal (2017); Lakshminarayanan et al. (2017); Meinshausen (2006).

2.4.2 The FLUXCOM Initiative. The FLUXCOM project represents a major advance in empirical flux up-scaling. Jung *et al.* synthesised machine-learning techniques to combine FLUXNET tower measurements with MODIS remote sensing and meteorological drivers, producing global gridded estimates of carbon and energy fluxes Jung et al. (2019, 2020). FLUXCOM developed two complementary configurations. The remote-sensing (RS) setup estimates fluxes using only MODIS inputs and produces 8-daily outputs at 0.0833° spatial resolution. The RS + METEO setup combines daily meteorological data with mean seasonal cycles of MODIS predictors to generate daily outputs at 0.5° resolution. A factorial design explored nine machine-learning methods for the RS configuration and three methods for the RS + METEO configuration, combined with multiple climate forcing data sets and two flux-partitioning schemes, yielding an ensemble of 120 global products Tramontana et al. (2016). Cross-validation showed good performance for net radiation and turbulent energy fluxes but larger errors for carbon fluxes, especially in regions with sparse tower coverage Jung et al. (2020). FLUXCOM quantifies uncertainty associated with algorithm choice, yet individual models assume unimodal predictive

distributions and provide only deterministic outputs, limiting interpretability for risk-sensitive applications.

2.4.3 Limitations and Motivation. Despite substantial progress, data-driven up-scaling faces several challenges. **(1) Spatial and temporal gaps**: tower measurements are unevenly distributed and often located in homogeneous or accessible ecosystems, leading to extrapolation into under-represented regions. **(2) Footprint mismatch**: dynamic footprints complicate alignment of tower fluxes with remote-sensing pixels and model grids. **(3) Deterministic predictions**: most machine-learning models output single values, neglecting the variability inherent in flux observations. **(4) Uncertainty characterization**: ensemble approaches provide sample variance across algorithms but cannot distinguish measurement noise from structural multimodality or epistemic uncertainty due to limited data Depeweg et al. (2018); Jung et al. (2020). These limitations motivate probabilistic models that generate full predictive distributions and decompose uncertainty into distinct components.

2.5 Probabilistic Regression and Calibration

Probabilistic regression models aim to estimate the entire conditional distribution $p(y \mid \mathbf{x})$ rather than a single point estimate. Given predictors $\mathbf{x}$ and a continuous target y, such models output a distribution $\hat{p}(y \mid \mathbf{x})$ that reflects predictive uncertainty. Key scoring rules assess distribution quality: (1) the *negative log-likelihood* (NLL) measures the average log probability assigned to observations; lower NLL indicates better fit and penalises over-confident distributions Gneiting and Raftery (2007); (2) the *continuous ranked probability score* (CRPS) integrates the squared difference between the predictive cumulative distribution and the Heaviside step function of the observation, balancing calibration and sharpness Gneiting, Balabdaoui, and Raftery (2007); and (3) the *probability integral transform* (PIT) and reliability diagrams evaluate calibration by comparing predicted quantile coverage with empirical frequencies; well-calibrated models yield a uniform PIT and coverage curves close to the 1:1 line Bröcker and Smith (2007); Gneiting et al. (2007); Gneiting and Raftery (2007). Proper scoring rules such as the Brier score and quantile loss are also used for discrete outcomes and quantiles. Calibration is crucial for risk-sensitive decisions in carbon markets and climate policy, where under- or over-confident predictive distributions can mislead carbon credit valuation or ecosystem management strategies. In addition, formal tests such as Diebold–Mariano statistics enable

comparison of predictive accuracy between competing probabilistic models Diebold and Mariano (1995); Kuleshov, Fenner, and Ermon (2018).

2.6 Mixture Models and Mixture Density Networks

2.6.1 Mixture of Gaussians Representation. Complex regression problems often exhibit multi-valued mappings from inputs to outputs. For example, a particular combination of vegetation index and temperature may correspond to different NEE values depending on unobserved factors such as soil moisture, phenological stage or disturbance history. Standard neural networks trained with mean squared error approximate the conditional mean, which can lie between modes and thus provide implausible predictions. To address this, mixture models represent the conditional density as a weighted sum of components:

$$p(y \mid \mathbf{x}) = \sum_{k=1}^{K} \pi_k(\mathbf{x})\, \mathcal{N}\big(y; \mu_k(\mathbf{x}),\, \sigma_k^2(\mathbf{x})\big),$$

where K is the number of components; $\pi_k(\mathbf{x})$ are non-negative mixture weights that sum to one; and $\mu_k(\mathbf{x})$, $\sigma_k^2(\mathbf{x})$ are component means and variances. Mixture models can capture multimodality and heteroscedasticity but require a mechanism to map inputs to distribution parameters.

2.6.2 Mixture Density Networks. Mixture density networks (MDN), introduced by Bishop Bishop (1994), combine a neural network with a mixture model. Given inputs $\mathbf{x}$, the network outputs the mixture weights, means and variances of a Gaussian mixture. Training maximizes the likelihood of the observed data (equivalently minimizing NLL). Because MDNs model the full conditional density, they can represent multi-valued relationships and provide heteroscedastic uncertainty estimates Bishop (1994). MDNs have been applied to inverse kinematics, speech synthesis and time-series forecasting. However, they capture only *aleatoric* (data) uncertainty: weights are deterministic parameters learned by maximum likelihood, so MDNs do not quantify *epistemic* uncertainty arising from limited data or model structure.

2.7 Bayesian Neural Networks

Bayesian neural networks (BNN) extend standard networks by placing probability distributions over weights and biases. Given a prior distribution $p(\mathbf{w})$ on network weights $\mathbf{w}$ and observed data $\mathcal{D}$, Bayes' rule yields the posterior $p(\mathbf{w} \mid \mathcal{D}) \propto p(\mathcal{D} \mid \mathbf{w})\, p(\mathbf{w})$. The predictive distribution for a new input $\mathbf{x}$ marginalises over

weight uncertainty:

$$p(y \mid \mathbf{x}, \mathcal{D}) = \int p(y \mid \mathbf{x}, \mathbf{w})\, p(\mathbf{w} \mid \mathcal{D})\, \mathrm{d}\mathbf{w}.$$

This formulation captures *epistemic* uncertainty due to lack of knowledge about the true parameter values. Exact inference is intractable for deep networks, so approximate methods such as variational inference, Monte Carlo dropout and deep ensembles are employed. Variational inference posits a parametric posterior family and minimizes the Kullback–Leibler divergence to the true posterior but may yield biased uncertainty estimates; Monte Carlo dropout interprets dropout masks at test time as approximate Bayesian sampling; deep ensembles average predictions from independently trained networks with different initializations. These methods capture some epistemic uncertainty but generally assume simple likelihoods (e.g., Gaussian) and cannot represent multimodal or skewed distributions Kendall and Gal (2017); Lakshminarayanan et al. (2017).

2.8 Summary and Research Gap

Estimating carbon fluxes at landscape to global scales requires combining sparse tower observations with multi-source environmental data and modeling complex, nonlinear, multivalued relationships. Traditional machine-learning models have improved flux up-scaling but provide deterministic predictions and limited uncertainty quantification. Ensemble approaches, such as FLUXCOM, consider algorithmic and input-data variability yet assume unimodal outputs and cannot disentangle aleatoric, compositional and epistemic uncertainty Jung et al. (2020). Mixture density networks capture multimodality but ignore weight uncertainty; Bayesian neural networks quantify epistemic uncertainty but assume simple likelihoods. Recent theoretical and practical advances suggest that hybrid architectures, such as Bayesian mixture density networks, could provide richer uncertainty characterization by modeling multimodal conditional densities and decomposing uncertainty. The next chapter presents our proposed methodology to address these gaps.

CHAPTER III
THEORY OF THREE-CATEGORY UNCERTAINTY

3.1 Motivation, definitions, and identifiability

Given an input x, we decompose predictive dispersion for a target Y into three terms Depeweg et al. (2018); Kendall and Gal (2017):

- **Aleatoric uncertainty** $A(x)$: variability that persists within a single mode of $p(y \mid x)$ due to the data generating process. This does not vanish with more data.
- **Compositional uncertainty** $C(x)$: variability across distinct modes of $p(y \mid x)$ that arises when inputs miss a disambiguating feature. Adding such a feature collapses $C(x)$.
- **Epistemic uncertainty** $E(x)$: variability of the model's predictive mean at x due to parameter or model uncertainty under finite data. With a correct model and dense data near x, $E(x)$ shrinks toward zero.

We refer to this three-component split as the ACE (aleatoric, compositional and epistemic) decomposition. Mathematically, ACE is a structured application of the law of total variance to multivalued problems with a latent "branch" variable; the novelty lies less in new probability theory than in making the within-branch and between-branch pieces explicit, naming the compositional term, and showing how to estimate all three components with mixture models in practice. Two guiding principles structure this chapter. First, $A(x)$ and $C(x)$ can be defined from the conditional law $p(y \mid x)$ itself, hence they admit model free ground truth in a suitable controlled construction. Second, $E(x)$ is inherently model dependent, so we define it with respect to a Bayesian posterior and treat it as a property of a chosen model class and prior rather than of nature. We provide an oracle reference that lower bounds practical $E(x)$ and isolates parameter uncertainty from component assignment uncertainty.

3.2 Controlled toy problem

We adopt a smooth, non-monotone forward map so that the inverse has multiple branches on a subset of the domain. Let

$$Y \sim \mathrm{Unif}(0,1), \qquad X = g(Y) + \eta, \qquad g(y) = y + a\sin(2\pi y), \qquad \eta \sim \mathrm{Unif}[-\delta, \delta], \tag{3.1}$$

with amplitude $a > 0$ and noise half width $\delta > 0$. Then

$$g'(y) = 1 + 2\pi a \cos(2\pi y),$$

so the map is globally non-monotone if and only if $a \geq 1/(2\pi)$. In that case there are two stationary points $y_1 < y_2$ that partition the unit interval into three monotone branches

$$B_1 = [0, y_1], \qquad B_2 = [y_1, y_2], \qquad B_3 = [y_2, 1].$$

By the inverse function theorem, each restriction $g_j = g \restriction_{B_j}$ is bijective onto its image and has a continuous inverse g_j^{-1}.

Notation. We fix the following objects used throughout:

- Feasible set at input x: $S_x = \{\, y \in [0,1] : |g(y) - x| \leq \delta \,\}$.
- Branch feasible intervals $I_j(x) = S_x \cap B_j = [\ell_j(x), r_j(x)]$, which can be empty.
- Lengths $L_j(x) = r_j(x) - \ell_j(x)$, and mixture weights $w_j(x)$ defined below.
- For a uniform law on $[\ell_j, r_j]$, mean $\mu_j(x) = \frac{1}{2}(\ell_j + r_j)$ and variance $v_j(x) = L_j(x)^2/12$.
- Mixture mean $\mu(x) = \sum_j w_j(x)\mu_j(x)$.

End points on each branch. Let $B_j = [b_j^{\mathrm{L}}, b_j^{\mathrm{R}}]$. The preimage on branch j of the measurement band $[x-\delta,\, x+\delta]$ is an interval after inversion and clamping to B_j. If g_j is increasing, set

$$\ell_j(x) = \max\{b_j^{\mathrm{L}},\, g_j^{-1}(x-\delta)\}, \qquad r_j(x) = \min\{b_j^{\mathrm{R}},\, g_j^{-1}(x+\delta)\}.$$

If g_j is decreasing, reverse the endpoints:

$$\ell_j(x) = \max\left\{b_j^{\mathrm{L}},\, \min\big(g_j^{-1}(x-\delta), g_j^{-1}(x+\delta)\big)\right\},$$
$$r_j(x) = \min\left\{b_j^{\mathrm{R}},\, \max\big(g_j^{-1}(x-\delta), g_j^{-1}(x+\delta)\big)\right\}.$$

These formulas yield $I_j(x) = [\ell_j(x), r_j(x)]$ and lengths $L_j(x) \geq 0$.

3.3 Geometry and mixture structure at fixed x

Under (3.1) with uniform prior and uniform noise, the conditional density is constant on the feasible set. The following proposition summarizes the structure.

Proposition 3.1 (Mixture of uniforms on branch intervals)**.** *Fix x. Let $J(x)$ be the number of non-empty branch intervals $I_j(x)$. Then*

$$Y \mid X = x \sim \sum_{j=1}^{J(x)} w_j(x)\, \mathrm{Unif}(\ell_j(x), r_j(x)), \qquad w_j(x) = \frac{L_j(x)}{\sum_{i=1}^{J(x)} L_i(x)}.$$

Moreover, for small δ and x in the interior of the image of branch j, $L_j(x) \approx 2\delta \,/\, |g'(y_j^\star)|$, where $y_j^\star$ solves $g(y_j^\star) = x$ on branch j.

A proof is provided in Appendix A, together with generalizations for non-uniform priors or noise.

3.4 Closed-form expressions for $A(x)$ and $C(x)$

Within branch j we have a uniform law on $[\ell_j, r_j]$, hence $v_j(x) = L_j(x)^2/12$ and $\mu_j(x) = (\ell_j + r_j)/2$. The mixture mean is $\mu(x) = \sum_j w_j \mu_j$. Define

$$A(x) = \sum_{j=1}^{J(x)} w_j(x)\, v_j(x) = \frac{1}{12} \sum_{j=1}^{J(x)} w_j(x)\, L_j(x)^2, \tag{3.2}$$

$$C(x) = \sum_{j=1}^{J(x)} w_j(x)\, \big(\mu_j(x) - \mu(x)\big)^2. \tag{3.3}$$

By construction, $A(x)$ is the within-branch variance averaged over branches, while $C(x)$ is the between-branch variance of the branch means around the mixture mean.

Proposition 3.2 (Exact variance decomposition in the controlled setting)**.** *Under the construction in* (3.1)*, for every x with at least one non-empty branch interval,*

$$\mathrm{Var}[Y \mid X = x] \;=\; A(x) + C(x).$$

The statement follows from the law of total variance for finite mixtures Depeweg et al. (2018). A proof is given in Appendix A.

Small δ asymptotics. When x lies well inside the image of branch j and δ is small, $L_j(x) \approx 2\delta \,/\, |g'(y_j^\star)|$ with $g(y_j^\star) = x$. In single valued regions of the inverse, $J(x) = 1$ hence $C(x) = 0$ and

$$A(x) \approx \frac{\delta^2}{3}\, |g'(y^\star)|^{-2}.$$

Beyond uniform priors or noise. If Y has density p_Y that is not uniform, or η has density r that is not uniform, then $p(y \mid x) \propto p_Y(y)\, r\big(x - g(y)\big)$ on the feasible set. The mixture-of-uniforms form no longer holds. It is still correct to define

$$A(x) = \sum_j w_j(x)\, \mathrm{Var}\big[\, Y \mid X = x, Z = j \,\big]$$

and

$$C(x) = \sum_j w_j(x)\, \big(\mathbb{E}[Y \mid X = x, Z = j] - \mathbb{E}[Y \mid X = x]\big)^2$$

where Z is a latent branch label and $w_j(x) = \mathbb{P}(Z = j \mid X = x)$. Under uniform assumptions these reduce to (3.2) and (3.3).

3.5 Epistemic uncertainty and an oracle reference

We define epistemic uncertainty at x with respect to a Bayesian posterior over parameters θ, following Bayesian deep-learning treatments of epistemic variance Depeweg et al. (2018); Kendall and Gal (2017):

$$E(x) \;=\; \mathrm{Var}_{\theta\sim p(\theta|\mathcal{D})}\big(\,\mathbb{E}[\,Y \mid x,\theta\,]\,\big),$$

where $\mathcal{D}$ is the training data. This captures dispersion in the predictive mean due to parameter uncertainty.

Unlike $A(x)$ and $C(x)$, which are defined directly from the conditional law $p(y \mid x)$, the value of $E(x)$ depends on the assumed model family, prior, and approximate posterior. Different Bayesian models that fit the data equally well can induce different epistemic curves. In later chapters we therefore interpret E comparatively across models, regions, or hyperparameters rather than as an absolute, model-free physical quantity.

Oracle with known branch identity. Introduce a branch label $Z \in \{1,2,3\}$. If the branch is known, fit per branch k a Bayesian regressor with features $\phi(x) = [1, x, \ldots, x^d]^\top$, Gaussian prior $\theta_k \sim \mathcal{N}(\mu_{0,k}, \Sigma_{0,k})$, and homoscedastic Gaussian observation noise. The posterior is Gaussian with covariance Σ_k^{post}. The variance of the predictive mean on branch k is

$$v_{\mathrm{epi},k}(x) \;=\; \phi(x)^\top \Sigma_k^{\mathrm{post}} \phi(x). \tag{3.4}$$

Define the *oracle* epistemic uncertainty by averaging over known branch identity,

$$E_{\mathrm{oracle}}(x) \;=\; \sum_k \tilde{w}_k(x)\, v_{\mathrm{epi},k}(x),$$

with weights $\tilde{w}_k(x)$ equal to the true branch posterior probabilities under the data generating process. This oracle removes component assignment uncertainty.

Proposition 3.3 (Practical epistemic exceeds the oracle)**.** *Let* $m_{\theta,Z}(x) = \mathbb{E}[Y \mid x, \theta, Z]$. *Then*

$$\mathrm{Var}_{\theta,Z}\big(m_{\theta,Z}(x)\big) \;\geq\; \mathbb{E}_Z\Big[\,\mathrm{Var}_\theta\big(m_{\theta,Z}(x)\big)\,\Big].$$

In particular, the practical epistemic variance of the predictive mean (marginalizing unknown component identity) lower-bounds the oracle variance that conditions on the true component.

CHAPTER IV

METHODOLOGY: BAYESIAN MIXTURE DENSITY NETWORKS FOR UQ

This chapter turns the theory in Chapter III into an operational estimator. We specify the Bayesian Mixture Density Network (BMDN) architecture, the training objective and approximate Bayesian inference, the exact mapping from model outputs to the three uncertainty terms $A(x)$, $C(x)$, $E(x)$, and the diagnostic toolkit for calibration and proper scoring Bröcker and Smith (2007); Gneiting et al. (2007); Gneiting and Raftery (2007); Kuleshov et al. (2018).

4.1 Overview

This chapter introduces a concrete neural architecture that is explicitly designed to approximate ACE decomposition in practice: the Bayesian Mixture Density Network (BMDN). A BMDN combines three ingredients:

1. A deterministic feature extractor that maps raw inputs into a learned representation.
2. One or more Bayesian layers with weight posteriors, which are the source of epistemic uncertainty.
3. A mixture density output head that predicts the parameters of a Gaussian mixture and thereby encodes aleatoric and compositional uncertainty.

The goals of this chapter are to:

- motivate why this hybrid architecture is appropriate for three-way uncertainty,
- specify the parameterization and priors used in the Bayesian layers,
- define the training objective, based on a variational evidence lower bound (ELBO),
- describe how to perform prediction by Monte Carlo sampling from the weight posterior, and
- show exactly how to map BMDN outputs into operational estimators $\hat{A}(x)$, $\hat{C}(x)$, and $\hat{E}(x)$ that correspond to the theoretical quantities from Chapter III.

In later chapters, this architecture will be instantiated for both the controlled problem and the carbon flux task, with task-specific hyperparameters and ablations.

4.2 Why a Bayesian Mixture Density Network?

4.2.1 Requirements implied by the A/C/E theory. The three-way decomposition in Chapter III imposes concrete structural requirements on any model that aims to approximate $A(x)$, $C(x)$, and $E(x)$:

- To represent compositional uncertainty $C(x)$, the model must have an explicit or implicit notion of multiple components z at fixed input x, with associated mixing weights and component-specific predictions.
- To represent aleatoric uncertainty $A(x)$, the model must produce within-component variances that can be interpreted as noise or micro-scale variability.
- To represent epistemic uncertainty $E(x)$, the model must have stochastic parameters θ whose posterior spread can be propagated into the predictive mean.

A standard deterministic neural network fails on all three counts: it has neither explicit mixture structure nor parameter uncertainty. A classical Mixture Density Network (MDN) addresses the first two requirements by outputting a conditional Gaussian mixture, but it treats all weights as point estimates and therefore cannot separate parameter uncertainty from compositional effects Bishop (1994, 2006); Hjorth and Nabney (1999). A fully Bayesian neural network satisfies the third requirement but does not, by itself, separate multimodality from noise unless an explicit mixture output is added Blundell et al. (2015); Gal and Ghahramani (2016); Lakshminarayanan et al. (2017); Neal (1996).

4.2.2 Hybrid architecture: deterministic backbone, Bayesian layers, mixture head. The BMDN is a hybrid that aligns naturally with the three-way framework:

- **Mixture head.** Given a representation $h(x)$, the MDN head outputs mixture weights, means, and variances of a Gaussian mixture for $Y \mid X = x$. This provides the latent components and their within-component variances required for $A(x)$ and $C(x)$.
- **Bayesian layers.** A subset of layers is treated as Bayesian, with a learned posterior over weights. Epistemic uncertainty $E(x)$ is then computed as the variance of the predictive mean across posterior samples of those weights.
- **Deterministic backbone.** The remaining layers form a deterministic feature extractor that learns expressive features from high-dimensional inputs (for example, carbon flux drivers). Keeping most of the network deterministic keeps training stable and computationally feasible.

This mirrors the logic of Bayesian Last Layer (BLL) architectures, which argue that a full Bayesian treatment of all layers is often unnecessary. A small Bayesian head on top of a deterministic encoder can capture most of the useful epistemic structure

at a fraction of the computational cost. The BMDN extends this idea from single Gaussians to mixtures, precisely to address compositional uncertainty in addition to aleatoric and epistemic components.

4.3 Architecture

We now describe the architecture in more concrete terms. The same template is used for the controlled problem and the carbon flux task, with only minor changes in depth, width, and input preprocessing.

Let $x \in \mathbb{R}^d$ denote the input features and $y \in \mathbb{R}$ the scalar target.

4.3.1 Inputs and preprocessing. For the synthetic problem, x consists of low-dimensional continuous features generated by the controlled data model in Section 3.2. For the carbon flux task, x is a concatenation of multispectral satellite bands and environmental drivers such as temperature, radiation, wind, humidity, and ancillary variables.

All continuous features are standardized using training-set statistics (zero mean and unit variance along each feature dimension). The target y is also standardized for training. Reported metrics and plots in subsequent chapters are expressed in the original physical units by applying the inverse transformation.

4.3.2 Deterministic feature extractor. The first part of the network is a deterministic feature extractor,

$$h_0(x) = x, \quad h_\ell(x) = \phi\big(W_\ell h_{\ell-1}(x) + b_\ell\big), \quad \ell = 1, \dots, L_{\text{det}},$$

where ϕ is a nonlinear activation such as ReLU or GELU, and W_ℓ, b_ℓ are standard trainable weights and biases.

For the synthetic problem, the feature extractor can be shallow, since the input is low-dimensional. For the carbon flux task, the extractor must integrate heterogeneous drivers into a compact representation, so a deeper multilayer perceptron with width W and optional pre-activation layer normalization is used. The deterministic backbone is trained jointly with the Bayesian layers but its parameters are always treated as point estimates.

4.3.3 Bayesian layer block. Above the deterministic backbone we introduce one or more Bayesian linear layers, following variational Bayes treatments of Bayesian neural networks Blundell et al. (2015); Kingma and Welling (2014). For clarity, consider the case of a single Bayesian layer:

$$z(x; \omega) = W_{\text{B}} h_{L_{\text{det}}}(x) + b_{\text{B}},$$

where $\omega = (W_{\mathrm{B}}, b_{\mathrm{B}})$ denotes the Bayesian weights. Instead of treating ω as a fixed vector, we place a Gaussian prior,

$$p(\omega) = \mathcal{N}\big(\omega \mid 0, \sigma^2_{\text{prior}} I\big),$$

and use variational inference to approximate the posterior $p(\omega \mid \mathcal{D})$ with a tractable family $q_\phi(\omega)$, typically a fully factorized Gaussian.

In variants with multiple Bayesian layers, the Bayesian block comprises a small subnetwork with all weights treated as random and endowed with similar Gaussian priors. The number of Bayesian layers and their width is an architectural degree of freedom that is examined empirically in the synthetic and real-data ablation studies in Chapters V and VI.

4.3.4 Mixture density output head. Given the Bayesian latent representation $z(x;\omega)$, the MDN head predicts the parameters of a K-component Gaussian mixture:

$$p(y \mid x, \omega) = \sum_{k=1}^{K} \pi_k(x;\omega)\, \mathcal{N}\big(y \mid \mu_k(x;\omega), \sigma_k^2(x;\omega)\big).$$

The head is implemented as three parallel affine maps from $z(x;\omega)$:

– **Mixture weights.**

$$a_k(x;\omega) = w_k^{(\pi)\top} z(x;\omega) + c_k^{(\pi)}, \quad \pi_k(x;\omega) = \frac{\exp(a_k(x;\omega))}{\sum_{j=1}^{K} \exp(a_j(x;\omega))}.$$

– **Component means.**

$$\tilde{\mu}_k(x;\omega) = w_k^{(\mu)\top} z(x;\omega) + c_k^{(\mu)}, \quad \mu_k(x;\omega) = \tilde{\mu}_k(x;\omega).$$

– **Component variances.**

$$s_k(x;\omega) = w_k^{(\sigma)\top} z(x;\omega) + c_k^{(\sigma)}, \quad \sigma_k^2(x;\omega) = \sigma^2_{\min} + \text{softplus}\,\big(s_k(x;\omega)\big),$$

where $\sigma^2_{\min} > 0$ is a small constant that enforces a minimum variance.

This parameterization guarantees that the mixture weights are non-negative and sum to one, and that all component variances are strictly positive. In both the synthetic-problem study and the carbon flux experiments, this structure proved numerically stable and expressive.

4.4 Training objective and variational inference

4.4.1 Likelihood and negative log likelihood. Given data $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^{N}$, the MDN predictive density at a fixed parameter draw (that is, for

a fixed ω) is

$$p(y_n \mid x_n; \omega) = \sum_{k=1}^{K} \pi_k(x_n; \omega)\, \mathcal{N}\big(y_n \mid \mu_k(x_n; \omega), \sigma_k^2(x_n; \omega)\big).$$

The corresponding negative log likelihood (NLL) is

$$\mathcal{L}_{\mathrm{NLL}}(\omega) = -\sum_{n=1}^{N} \log \left[\sum_{k=1}^{K} \pi_k(x_n; \omega)\, \mathcal{N}\big(y_n \mid \mu_k(x_n; \omega), \sigma_k^2(x_n; \omega)\big) \right].$$

For numerical stability, the log mixture density is evaluated using the standard log-sum-exp trick, that is, by factoring out $\max_k z_{nk}$ where z_{nk} denotes the log component density plus log weight.

4.4.2 Variational objective. Let W denote all Bayesian weights and ϕ the variational parameters. The variational objective is the negative evidence lower bound (ELBO),

$$\mathcal{L}_{\mathrm{ELBO}}(\phi) = \mathbb{E}_{q_\phi(W)}\big[\mathcal{L}_{\mathrm{NLL}}(W)\big] + \beta\, \mathrm{KL}\big(q_\phi(W) \,\|\, p(W)\big),$$

where β is a scalar that can be ramped from 0 to 1 during training (KL annealing) to stabilize optimization in the early phases.

Exact hyperparameter settings and variants are documented in the experimental chapters.

4.4.3 Variational family and reparameterization. The variational family is chosen as a fully factorized Gaussian,

$$q_\phi(W) = \mathcal{N}\big(W \mid m, S\big),$$

where m is the mean and S is diagonal. The reparameterization trick writes samples as

$$W^{(s)} = m + S^{1/2} \epsilon^{(s)}, \quad \epsilon^{(s)} \sim \mathcal{N}(0, I),$$

which allows gradients to flow through the sampled weights. The expectation over $q_\phi(W)$ in the ELBO is approximated by Monte Carlo using a small number of such samples per mini-batch.

4.4.4 Practical optimization. All models are trained with the AdamW optimizer, cosine learning-rate decay with warmup, and gradient norm clipping. We use KL annealing as described above. Early stopping is applied based on a combination of validation NLL and calibration metrics described in Section 4.7. For each experiment, we record random seeds, the number of mixture components K, the Bayesian depth (number of Bayesian layers), and the number of posterior

weight samples S. Full training configurations are provided in the appendix for reproducibility.

4.5 Prediction and posterior sampling

Once the BMDN is trained, we need to perform posterior predictive inference to obtain both point predictions and the components of uncertainty.

4.5.1 Posterior sampling procedure. Given a test input x, predictive inference proceeds as follows:

1. Draw S samples of Bayesian weights,
$$W^{(1)}, \ldots, W^{(S)} \sim q_\phi(W).$$
2. For each $W^{(s)}$, compute the mixture parameters
$$\left\{\pi_k^{(s)}(x),\ \mu_k^{(s)}(x),\ \sigma_k^{2,(s)}(x)\right\}_{k=1}^{K}.$$
3. For diagnostics that require full predictive distributions (such as PIT or CRPS), optionally sample $y^{(s,r)}$ from each mixture $p(y \mid x, W^{(s)})$.

From these quantities we can compute:

- the posterior predictive mean
$$\hat{\mu}(x) = \frac{1}{S}\sum_{s=1}^{S} m_s(x), \quad m_s(x) = \sum_{k=1}^{K} \pi_k^{(s)}(x)\,\mu_k^{(s)}(x),$$
- the predictive variance and its decomposition into $\hat{A}(x)$, $\hat{C}(x)$, and $\hat{E}(x)$, as described in Section 4.6.

The number of samples S is a tunable hyperparameter. Experiments in later chapters show that 50 samples are sufficient to stabilize $\hat{E}(x)$ while keeping computation manageable. For expensive diagnostics we can use more samples than for routine prediction.

4.5.2 Computational cost. Relative to a deterministic MDN, the main additional costs are:

- sampling from $q_\phi(W)$ at training and test time,
- evaluating the KL term and its gradients, and
- computing Monte Carlo averages over the sampled weights.

Because only a small subset of layers is Bayesian, these costs scale with the size of the Bayesian block, not with the full network. The total computational cost is approximately linear in:

- the number of mixture components K,
- the number of Bayesian weights $|W|$, and

– the number of posterior samples S.

This matches the efficiency arguments in the BLL literature: concentrating Bayesian treatment in the last layers retains good uncertainty properties while keeping training and prediction tractable.

4.6 Operational estimators for $A(x)$, $C(x)$, $E(x)$

Chapter III defined $A(x)$, $C(x)$, and $E(x)$ in terms of conditional expectations and variances over latent components and parameters. The BMDN provides direct empirical estimators of those quantities.

Everything in this section is exact for a Gaussian mixture at fixed parameters and follows directly from the law of total variance. The only approximation arises from Monte Carlo averaging over posterior draws of W.

4.6.1 Within one posterior draw. Fix a posterior sample $W^{(s)}$. For each input x, the mixture head yields

$$\left\{\pi_k^{(s)}(x),\ \mu_k^{(s)}(x),\ \sigma_k^{2,(s)}(x)\right\}_{k=1}^{K}.$$

Define the mixture mean

$$m_s(x) = \sum_{k=1}^{K} \pi_k^{(s)}(x)\, \mu_k^{(s)}(x).$$

Define the within- and between-component terms for that mixture as

$$A_s(x) = \sum_{k=1}^{K} \pi_k^{(s)}(x) \left(\sigma_k^{(s)}(x)\right)^2, \tag{4.1}$$

$$C_s(x) = \sum_{k=1}^{K} \pi_k^{(s)}(x) \left(\mu_k^{(s)}(x) - m_s(x)\right)^2. \tag{4.2}$$

For fixed parameters $W^{(s)}$, the law of total variance for the latent mixture index Z gives

$$\mathrm{Var}\left(Y \mid x, W^{(s)}\right) = A_s(x) + C_s(x).$$

Importantly, label permutations of mixture components do not affect $A_s(x)$ or $C_s(x)$, since both depend only on mixtures of means and variances weighted by $\pi_k^{(s)}(x)$.

4.6.2 Aggregating across posterior draws. Let

$$\bar{m}(x) = \frac{1}{S} \sum_{s=1}^{S} m_s(x)$$

denote the Monte Carlo estimate of the posterior predictive mean. We then define

$$\widehat{A}(x) = \frac{1}{S}\sum_{s=1}^{S} A_s(x), \quad \widehat{C}(x) = \frac{1}{S}\sum_{s=1}^{S} C_s(x),$$

as Monte Carlo estimators of the posterior expectations $\mathbb{E}_W[A(x;W)]$ and $\mathbb{E}_W[C(x;W)]$, respectively.

To estimate epistemic uncertainty, we use the spread of the mixture mean across posterior draws:

$$\widehat{E}(x) = \frac{1}{S}\sum_{s=1}^{S} \big(m_s(x) - \bar{m}(x)\big)^2.$$

This is the sample variance of the mixture means and corresponds to the parameter-level variance $\mathrm{Var}_W[\mathbb{E}(Y \mid x, W)]$ from Chapter III.

By the law of total variance across the posterior,

$$\mathrm{Var}\left(Y \mid x, \mathcal{D}\right) \approx \widehat{A}(x) + \widehat{C}(x) + \widehat{E}(x),$$

with equality in the limit of infinitely many posterior samples and infinitely many data points. In practice, S is finite, so $\widehat{E}(x)$ has Monte Carlo noise. This can be reduced by increasing S for diagnostic runs.

Numerical implementations use stable formulas for variance and double-precision accumulators where necessary to avoid cancellation when values are close.

Algorithm 1 ACE estimation from a trained BMDN at input x

1: Sample $W^{(1)}, \ldots, W^{(S)} \sim q_\phi(W)$.
2: **for** $s = 1, \ldots, S$ **do**
3: Compute mixture parameters $\{\pi_k^{(s)}(x), \mu_k^{(s)}(x), \sigma_k^{2,(s)}(x)\}_{k=1}^K$.
4: Compute mixture mean $m_s(x) = \sum_k \pi_k^{(s)}(x)\, \mu_k^{(s)}(x)$.
5: Compute within-component term $A_s(x) = \sum_k \pi_k^{(s)}(x)\, \sigma_k^{2,(s)}(x)$.
6: Compute between-component term $C_s(x) = \sum_k \pi_k^{(s)}(x) \left(\mu_k^{(s)}(x) - m_s(x)\right)^2$.
7: **end for**
8: Set $\bar{m}(x) = \frac{1}{S}\sum_s m_s(x)$.
9: Set $\widehat{A}(x) = \frac{1}{S}\sum_s A_s(x)$, $\widehat{C}(x) = \frac{1}{S}\sum_s C_s(x)$.
10: Set $\widehat{E}(x) = \frac{1}{S}\sum_s \big(m_s(x) - \bar{m}(x)\big)^2$.

4.7 Calibration diagnostics

Beyond point-wise decomposition into $A(x)$, $C(x)$, and $E(x)$, we require that the full predictive distributions from the BMDN be well calibrated. Calibration concepts and scoring rules (negative log-likelihood, CRPS, PIT and reliability curves)

are introduced in Chapter II, Section 2.5; here we briefly summarize how they are used operationally.

We monitor:

- prediction interval coverage for central intervals (for example, 50%, 80%, 90%) computed from the posterior predictive mixture and compared to nominal levels on held-out data;
- PIT histograms constructed from posterior predictive CDF values at observed targets, which should be close to uniform under perfect calibration;
- reliability curves based on empirical versus nominal coverage across quantile levels, together with proper scoring rules such as NLL and CRPS.

If substantial miscalibration is detected, we may consider simple post-hoc calibration strategies (such as temperature scaling or isotonic regression on predictive quantiles). In the experiments presented in later chapters, we primarily rely on the BMDN architecture and training objective to deliver well-calibrated predictive distributions and use post-hoc calibration only as a diagnostic comparison.

CHAPTER V
SYNTHETIC VALIDATION OF THE BAYESIAN MIXTURE DENSITY NETWORK

The goal of this chapter is to determine whether the Bayesian Mixture Density Network (BMDN) introduced in Chapter IV recovers the desired three–way uncertainty decomposition in a controlled setting. Specifically, we ask whether, given the theoretical framework from Chapter III, a practical BMDN can approximate the aleatoric, compositional, and epistemic components well enough that we can trust its outputs on the real carbon flux task.

To answer this question we work with a synthetic problem where the data generating process is fully known. This setting mirrors the structural difficulty of carbon flux prediction, but it is sufficiently simple that the oracle uncertainty components can be computed analytically or via a tractable Bayesian linear regression (BLR) oracle. This allows a detailed comparison between theory and practice.

5.1 Experimental protocol and metrics

5.1.1 Controlled synthetic problem and its link to carbon flux. We adopt the non–monotone forward map from Chapter III. The latent target variable Y is drawn from a uniform prior on an interval that excludes a small margin near 0 and 1, and is mapped to the observed covariate X through a smooth, S–shaped function with additive sensor noise:

$$Y \sim \mathrm{Unif}(\mathrm{margin}, 1 - \mathrm{margin}), \tag{5.1}$$

$$g(y) = y + 0.3 \sin(2\pi y), \tag{5.2}$$

$$X = g(Y) + \eta, \tag{5.3}$$

$$\eta \sim \mathrm{Unif}(-\delta_x, \delta_x), \qquad \delta_x \approx 0.05. \tag{5.4}$$

The forward map g is not monotone. Over part of its range, the inverse map g^{-1} has three distinct branches: for many values of x there are up to three plausible values of y that could have produced that observation. This multi–valued structure is exactly what gives rise to compositional uncertainty in Chapter III, because the conditional distribution $p(Y \mid X = x)$ is genuinely multimodal rather than merely noisy.

This toy problem acts as a proxy for the carbon flux task in two specific ways.

Multi–valued inverse structure. In the synthetic problem, the same value of x can correspond to three different values of y, depending on which branch of g^{-1} generated the sample. In the carbon flux setting, the same combination of environmental drivers and spectral reflectances can plausibly correspond to multiple net ecosystem exchange (NEE) values, for example due to unobserved soil properties, canopy structure, or microbial activity. Conceptually, the three branches of g^{-1} act as stand–ins for different "flux regimes".

Decomposition of predictive variance. In the synthetic problem we can decompose the conditional variance $\mathrm{Var}(Y \mid X = x)$ into aleatoric, compositional, and epistemic components, and compute reference values either analytically or using the BLR oracle from Chapter III. This mirrors our goal in the real carbon flux application, where we eventually want to interpret uncertainty patterns in terms of measurement noise, regime ambiguity, and model ignorance.

The synthetic experiments therefore serve as a stress test of identifiability for the three components $A(x)$, $C(x)$, and $E(x)$. They are designed to answer whether a practical BMDN can reproduce the theoretical behavior of the decomposition in a situation that has the same structural difficulty as carbon flux prediction but where the "correct" answers are known.

5.1.2 Data generation and splits.

We generate paired observations (x, y) by sampling y from the uniform prior in (5.1), applying the forward map (5.2)–(5.3), and adding uniform sensor noise (5.4) to obtain x. All samples lie in the region where the analytic inverse exists, so the closed–form oracle expressions for the aleatoric, compositional, and epistemic uncertainties from Chapter III remain valid.

Unless otherwise stated, we use

$$N_{\text{train}} = 8{,}000, \qquad N_{\text{val}} = 2{,}000, \qquad N_{\text{test}} = 2{,}000,$$

with each split drawn independently from the same process. There is no deliberate distribution shift between train, validation, and test. This design ensures that epistemic uncertainty $E(x)$ can genuinely shrink toward the oracle baseline as the training set grows, rather than being dominated by covariate shift.

Before training, both x and y are normalized to approximately $[-1, 1]$. All reported evaluation metrics are computed in the original units of y, by inverting the normalization on the model outputs.

5.1.3 Model definitions. We evaluate four models on this controlled task. Full architectural details and training procedures are given in Chapter IV; here we summarize the key distinctions.

M0: Heteroscedastic Gaussian regressor. Model M0 is a standard neural network that outputs, for each input x, the mean and variance of a Gaussian predictive distribution. This allows the model to represent input–dependent noise, but the predictive distribution is unimodal and the model does not provide any decomposition into aleatoric, compositional, and epistemic terms. M0 serves as a strong baseline for point prediction and for assessing the benefit of mixture and Bayesian structure.

M1: Mixture Density Network (MDN). Model M1 is a deterministic mixture density network with K Gaussian components. For each input x it outputs the component means, variances, and mixture weights. All network weights are treated as point estimates. M1 can represent multimodal conditional distributions and therefore has a meaningful within–component and between–component variance structure, but it has no Bayesian treatment of parameters.

M2: Bayesian Mixture Density Network (BMDN). Model M2 is the proposed Bayesian MDN. It shares the deterministic feature extractor and mixture head with M1, but the last feature layer is treated as Bayesian. A Gaussian prior is placed on its weights and the posterior is approximated with stochastic variational inference. This yields a posterior distribution over network parameters and therefore a nonzero epistemic component $E(x)$ that reflects parameter uncertainty, in addition to the aleatoric and compositional components derived from the mixture head.

M3: BMDN with disambiguating feature. Model M3 augments M2 with an auxiliary input feature z that encodes the latent branch index of g^{-1}. When z is perfectly informative, the conditional distribution $p(Y \mid X, Z)$ becomes essentially unimodal and compositional uncertainty should collapse. In practice we can also corrupt z with a configurable flip probability to simulate partially informative regime indicators. M3 is intentionally not realistic for carbon flux, since it uses label–derived information, but it is extremely useful as a diagnostic for how much compositional variance could be reduced by side information about regimes.

In the main model comparison we focus on M0, M1, and M2, because these models operate in the realistic setting where the network sees only x. Model M3 is used later in a targeted ablation.

5.1.4 Evaluation metrics and calibration diagnostics. We evaluate each model on the test set using both point prediction metrics and probabilistic diagnostics.

Point prediction metrics. For point prediction, we use the mean of the MDN predictive distribution (the mixture-averaged mean) as $\hat{y}(x)$. We report:

- Root mean squared error (RMSE) between the observed fluxes and this MDN predictive mean; RMSE is always non-negative and smaller values indicate better point prediction accuracy.
- Coefficient of determination R^2, computed using the same MDN predictive mean as the point prediction; R^2 takes values less than or equal to 1, with 1 indicating a perfect fit and values near or below 0 indicating performance comparable to, or worse than, predicting the sample mean.

Probabilistic scoring rules. We report:

- Negative log likelihood (NLL) under the predictive distribution, averaged over test examples; lower NLL means that the model assigns higher probability density to the observed fluxes and therefore fits the data better in a probabilistic sense.
- Continuous ranked probability score (CRPS), averaged over the test set; CRPS compares the full predictive distribution to the observed value, has the same units as the target variable, and smaller values indicate sharper and better aligned predictive distributions.

Calibration diagnostics. We assess calibration using:

- Reliability curves, which plot empirical coverage as a function of nominal central prediction interval level; a perfectly calibrated model lies on the diagonal, while systematic deviations reveal underconfident or overconfident intervals.
- Empirical coverage of 50%, 80%, and 90% central prediction intervals, where values close to the nominal levels indicate well calibrated uncertainty and large deviations signal miscalibration.

Table 1. Test set performance on the synthetic problem. Values are reported for root mean squared error (RMSE), coefficient of determination (R^2), negative log likelihood (NLL), continuous ranked probability score (CRPS), and empirical coverage of central 50%, 80%, and 90% prediction intervals.

Model	RMSE	R^2	NLL	CRPS	Cov_{50}	Cov_{80}	Cov_{90}
M0: Heteroscedastic Gaussian	0.185	0.487	-0.693	0.096	0.488	0.778	0.887
M1: MDN(K=3)	0.184	0.490	-1.540	0.088	0.466	0.780	0.890
M2: Bayesian MDN(K=3)	0.185	0.487	-1.519	0.088	0.462	0.784	0.896
M3: BMDN + branch feature*	0.030	0.987	-2.249	0.016	0.442	0.741	0.877

*M3 receives an auxiliary oracle branch indicator z constructed from labels, and is not available in realistic applications. It is included only as a diagnostic reference.

– Probability integral transform (PIT) histograms, which show the distribution of predictive CDF values evaluated at the observations; a uniform PIT histogram indicates calibrated predictive distributions, while U-shaped or skewed histograms indicate overdispersion, underdispersion, or bias.

Together, these metrics address two questions. First, whether the Bayesian last layer preserves the strong point prediction performance of a deterministic model. Second, whether it produces calibrated and interpretable uncertainty components whose behavior is consistent with the theoretical decomposition when the model or data are perturbed.

5.2 Overall model performance

Table 1 summarizes the test set metrics for M0, M1, and M2 on the synthetic problem. For completeness, M3 is also reported as a diagnostic oracle.

For the three realistic models M0, M1, and M2, RMSE and R^2 are essentially indistinguishable. In this regime with ample data and a smooth target, both a heteroscedastic Gaussian regressor and a deterministic MDN already learn a near–optimal conditional mean, so there is little room for improvement in point prediction.

The deterministic MDN M1 yields a substantial gain in probabilistic performance over M0. Its NLL is much lower (better) and CRPS improves from approximately 0.096 to 0.088. The Bayesian MDN M2 preserves these gains: its NLL is only slightly worse than M1, and its CRPS is effectively identical. This shows that the mixture structure is the main driver of probabilistic improvement, and the Bayesian last layer does not sacrifice calibration.

Interval coverage is slightly below nominal for all three models, which is typical for neural probabilistic regressors. Empirical coverage for 50% intervals lies between 0.46 and 0.49, and 80% and 90% intervals cover between approximately 0.78 and 0.90 of the test points. M2 achieves slightly better coverage at the 90% level than M0 and M1, while matching them at the central levels.

Model M3, which receives the oracle branch indicator, is almost perfectly accurate: it achieves RMSE of order 3×10^{-2}, $R^2 \approx 0.99$, excellent CRPS, and very favorable NLL. These numbers highlight the inherent difficulty of the original problem without regime information, and they provide an upper bound on what could be achieved if high–quality regime indicators were available in the carbon flux setting.

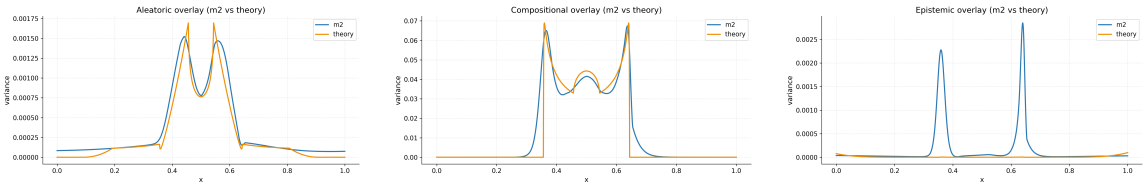
Overall, Table 1 shows that the Bayesian MDN M2 matches the strong point prediction performance of M0 and M1, retains the probabilistic gains of the MDN, and adds a meaningful epistemic component without degrading calibration.

5.3 Recovery of aleatoric, compositional, and epistemic uncertainty

The central validation question is whether the BMDN recovers the theoretical decomposition from Chapter III. To answer this, we evaluate M2 on a dense grid of x values and compute:

- The oracle curves $A^\star(x)$, $C^\star(x)$, and $E^\star(x)$ from the analytic or BLR–based reference.
- The empirical curves $A(x)$, $C(x)$, and $E(x)$ from the BMDN decomposition.

Figure 1 shows the overlay of these theoretical and empirical components.



(a) Aleatoric $A(x)$ (b) Compositional $C(x)$ (c) Epistemic $E(x)$

Figure 1. Recovered aleatoric, compositional, and epistemic uncertainty components from the Bayesian MDN on the synthetic problem. Each panel shows the BMDN estimate overlaid with the oracle curve from Chapter III.

Two observations stand out. First, in terms of shape, both the aleatoric and compositional curves closely track their oracle counterparts. Aleatoric uncertainty $A(x)$ is nearly flat across most of the domain, with a scale that matches the injected sensor noise and slight variations where the forward map g is locally steeper.

Compositional uncertainty $C(x)$ is almost zero in regions where the inverse map of g is effectively single valued, and it forms a broad peak in the central region where the inverse map is three valued. These patterns reproduce the theoretical structure derived in Chapter III.

Second, the magnitudes of the decomposed components are quantitatively aligned with the oracle. The mean values of $A(x)$ and $C(x)$ over the test grid agree with the oracle means to within a few percent, and the sum $A(x) + C(x) + E(x)$ matches the Monte Carlo predictive variance at each point, up to small numerical error. Epistemic uncertainty $E(x)$ has a shape similar to the BLR oracle but somewhat larger peaks in the strongly multimodal region. This is expected: the BLR oracle uses branch labels and a simpler within–branch model, while the BMDN must infer everything from x and shares parameters across modes, which leads to a broader posterior.

These results provide strong empirical support for the decomposition theorem from Chapter III: in a multi–valued problem with sufficient data, the within–component and between–component variances of a flexible MDN, combined with a Bayesian treatment of parameters, can approximate the true aleatoric, compositional, and epistemic components of $\mathrm{Var}(Y \mid X = x)$.

5.4 Model calibration

Calibration is critical if uncertainty estimates are to be used in downstream decisions. We therefore examine reliability curves, coverage of central prediction intervals, and PIT histograms for the main models. Figure 2 summarizes these diagnostics.

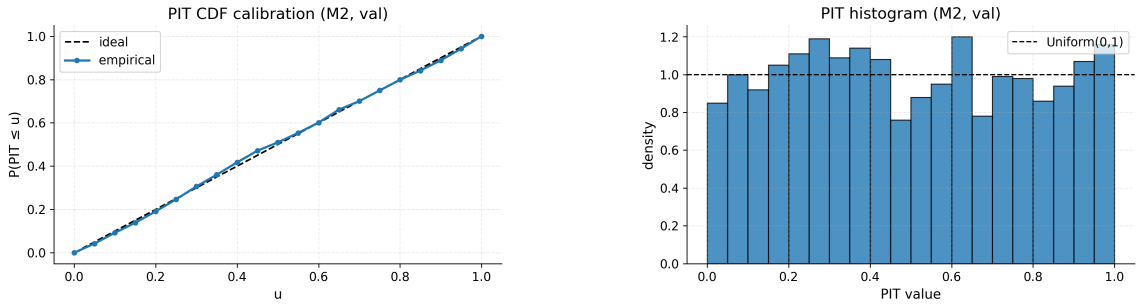


Figure 2. Calibration diagnostics. Left: PIT-based calibration (or reliability) curve for the CDF for M2 (BMDN). Right: PIT histograms for M2 (BMDN).

The BMDN model appears reasonably well calibrated. The PIT CDF calibration curve lies very close to the diagonal, with no evident S-shape, indicating no strong under- or over-dispersion. Coverage diagnostics show that the empirical coverages of the 50%, 80%, and 90% prediction intervals are essentially equal to their nominal levels, with only minor deviations attributable to sampling variability. PIT histograms are close to uniform, with only mild deviations, further supporting good overall calibration.

The deterministic MDN M1 achieves substantially better NLL and CRPS than the Gaussian regressor M0 while maintaining similar calibration. Relative to M0, both M1 and the Bayesian MDN M2 move the reliability curves closer to the diagonal and make the PIT histograms closer to uniform, indicating better-calibrated predictive distributions. The Bayesian MDN M2 preserves the probabilistic performance of M1 and yields calibration that is at least as good. In particular, its empirical coverage at the 90% level is slightly closer to nominal than that of M0 and M1, and its PIT histogram is closer to uniform in the tails.

Overall, the BMDN provides well-calibrated predictive distributions on the synthetic task, comparable to the best deterministic model, while adding an epistemic component that is not available in M0 or M1.

5.5 Ablation studies

To test the robustness and interpretability of the three–way decomposition, we perform a series of controlled ablations that modify the mixture capacity, the depth and prior of the Bayesian layer, the training set size, the availability of a disambiguating feature, and the level of data noise. In each ablation we keep as many factors fixed as possible and observe how the mean aleatoric, compositional, and epistemic variances, as well as the standard metrics, respond.

5.5.1 A1: Effect of the number of mixture components. The BMDN represents the conditional predictive distribution $p(y \mid x)$ as a Gaussian mixture with K components. In principle, K should be large enough to capture the multimodal structure of the target function, but not so large that the model becomes unnecessarily complex. In this ablation we vary only the number of components in the MDN head, sweeping $K \in \{1, \ldots, 9\}$, while keeping the deterministic feature extractor, Bayesian last layer, training data, optimizer, and priors identical.

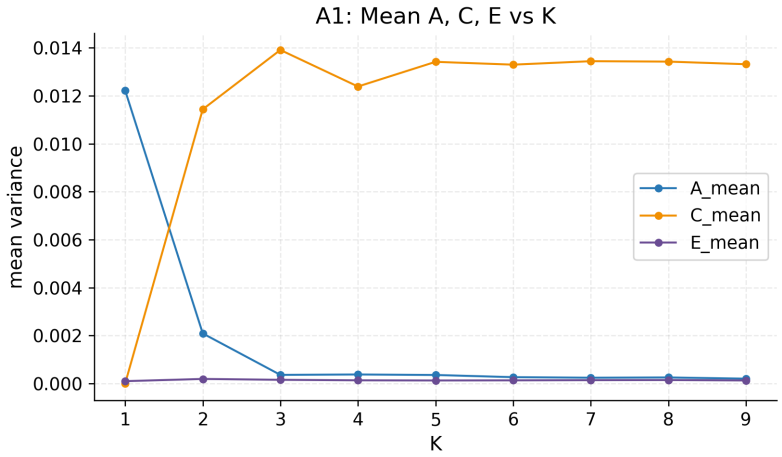


Figure 3. Ablation A1: mean aleatoric $A(x)$, compositional $C(x)$, and epistemic $E(x)$ variance as a function of the number of mixture components K in the MDN head. For each K, the means are computed over the validation set.

For each configuration we train the model to convergence and compute, on the validation set, the dataset–wide mean of each uncertainty component, $\mathbb{E}[A(x)]$, $\mathbb{E}[C(x)]$, $\mathbb{E}[E(x)]$, together with RMSE, NLL, CRPS, and R^2.

Figure 3 shows the effect of K on the uncertainty decomposition.

With $K = 1$, the model reduces to a single Gaussian per input and has no way to represent multiple branches. In this regime, the mean aleatoric variance is large and accounts for almost all of the total predictive variance, while the compositional term is identically zero. As soon as K increases to 2, the model can start assigning different branches to different components, the mean aleatoric variance drops by almost an order of magnitude, and the compositional variance becomes nonzero. For $K \geq 3$, the aleatoric term stabilizes at a very small value that matches the injected noise plus numerical error, while the compositional term saturates at a stable level. Epistemic variance remains small and nearly constant across the entire sweep.

Standard metrics show that RMSE and R^2 are essentially flat across K, while NLL and CRPS improve sharply moving from $K = 1$ to $K = 2$ or 3 and then exhibit a clear elbow. Once K is large enough to represent the three branches of the inverse map, additional components mainly introduce redundancy. Very large values of K lead to slightly worse NLL and marginally underdispersed predictive distributions, reflecting overfitting of small-scale structure in the likelihood. The main effect of K is

therefore on how predictive variance is allocated between aleatoric and compositional terms, rather than on point accuracy.

Based on this ablation, in all subsequent experiments we fix K to a value in the saturation regime where aleatoric variance has collapsed to a noise floor, compositional variance has stabilized, and epistemic variance remains small and decoupled from K.

5.5.2 A2: Bayesian depth in the feature extractor. The BMDN uses a deterministic feature extractor followed by a Bayesian layer and a deterministic MDN head. Epistemic uncertainty should arise primarily from uncertainty in the Bayesian part of the network. This ablation varies the *Bayesian depth* of the feature extractor to test how the decomposition responds.

Starting from the default architecture with three fully connected feature layers, we consider:

- `last`: two deterministic feature layers, one Bayesian layer, MDN head (default).
- `last_two`: one deterministic feature layer, two Bayesian layers, MDN head.
- `full`: three Bayesian feature layers, MDN head.

The MDN head remains deterministic in all cases. For each configuration we train with the same optimizer, prior scale, and schedule, then estimate the mean $A(x)$, $C(x)$, and $E(x)$ on the validation set and compute RMSE, R^2, NLL, and CRPS.

Figure 4 summarizes the mean uncertainties as a function of Bayesian depth.

Across all three configurations, compositional variance dominates the variance budget, with mean values around 1.3×10^{-2} to 1.5×10^{-2}. Aleatoric and epistemic components are an order of magnitude smaller. The aleatoric term changes only mildly with Bayesian depth, and the compositional term is similarly robust. In contrast, the epistemic term reacts strongly and monotonically: relative to `last`, the mean epistemic variance increases by approximately a factor of 3.3 for `last_two` and by about a factor of 4.2 for `full`. With three Bayesian layers, the epistemic component becomes larger than the aleatoric component on average.

Point prediction metrics are almost unchanged across the three configurations. The default `last` setting achieves the best NLL and CRPS, while deeper Bayesianization slightly degrades both scoring rules. The wider predictive distributions induced by multiple Bayesian layers are not rewarded by the proper scoring rules and appear overly diffuse relative to the data.

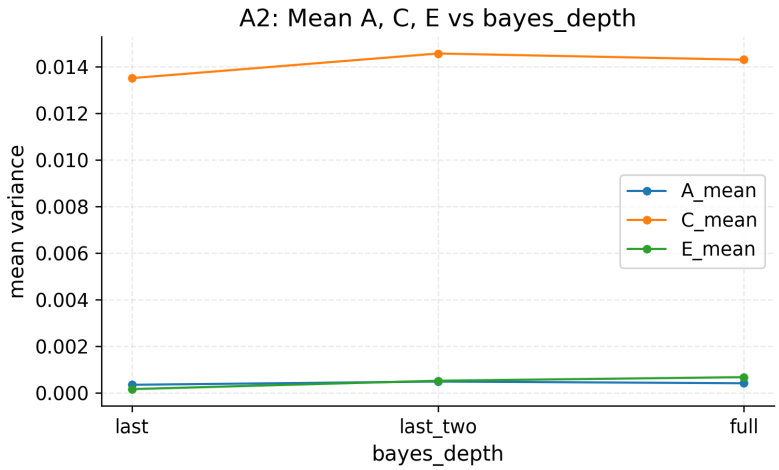
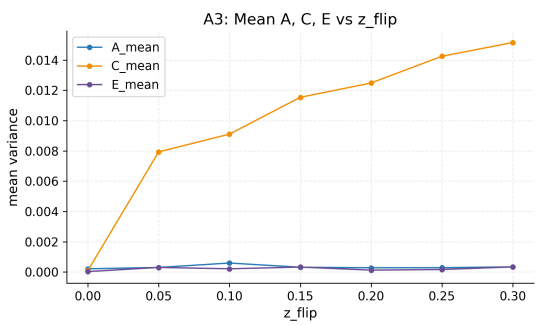


Figure 4. Ablation A2: mean aleatoric $A(x)$, compositional $C(x)$, and epistemic $E(x)$ variance for different Bayesian depths in the feature extractor. The three configurations correspond to making only the last, the last two, or all three feature layers Bayesian.

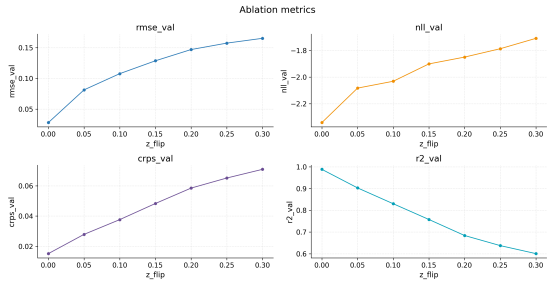
These results show that the decomposition responds in a principled way to Bayesian depth: increasing the number of Bayesian layers primarily inflates epistemic variance, while leaving aleatoric and compositional variance comparatively stable. From the perspective of calibration and computational cost, the default configuration with a single Bayesian layer provides the best trade–off, and this is the configuration used in the remaining experiments.

5.5.3 A3: Reducing compositional uncertainty with a disambiguating feature. The synthetic problem is structurally multimodal: for many values of x there are multiple plausible branches of y. In the baseline BMDN this ambiguity appears as compositional uncertainty $C(x)$, since the mixture components disagree about which branch is active. In this ablation we test whether a disambiguating feature can selectively collapse $C(x)$ and how the decomposition responds as that feature becomes noisy.

We augment the BMDN input with a scalar feature z that encodes the latent branch index. For each training example we compute the true branch $b \in \{1, \dots, B\}$ and then generate a potentially corrupted indicator z by flipping b to one of the other branches with probability $p_{\text{flip}} \in \{0.00, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30\}$. The case $p_{\text{flip}} = 0$ corresponds to a perfect disambiguator; larger values make z progressively less reliable. For each p_{flip} we train a fresh BMDN with the same architecture, priors,



(a) ACE vs p_{flip}



(b) Metrics vs p_{flip}

Figure 5. Ablation A3: effect of a disambiguating feature on the three–way decomposition. Left: mean aleatoric $A(x)$, compositional $C(x)$, and epistemic $E(x)$ as a function of the flip probability p_{flip} used to corrupt the branch indicator z. Right: RMSE, NLL and CRPS on the validation set as a function of p_{flip}.

and training schedule as in the main experiment, then evaluate predictive performance and the three uncertainty terms on a held–out grid.

Figure 5 shows how the mean uncertainties respond to the reliability of the disambiguating feature.

With a perfect disambiguator ($p_{\text{flip}} = 0$), the regression problem becomes almost deterministic. The BMDN achieves $R^2 \approx 0.99$, very low RMSE, and excellent probabilistic scores. In this regime, the mean compositional variance is smaller than the aleatoric term, indicating that once the branch is known the conditional distribution $p(y \mid x, z)$ is essentially unimodal.

Introducing even a small amount of noise in the disambiguator has a dramatic and highly selective effect on $C(x)$. At $p_{\text{flip}} = 0.05$, the mean compositional variance increases by roughly two orders of magnitude compared to the perfect case, and it already accounts for more than 90% of the total predictive variance. Aleatoric and epistemic components increase only modestly. As p_{flip} continues to grow, compositional variance becomes increasingly dominant, accounting for more than 95% of the total by $p_{\text{flip}} = 0.30$, while the scales of $A(x)$ and $E(x)$ remain bounded.

Predictive metrics degrade smoothly as the disambiguator becomes noisier: RMSE increases, R^2 decreases, and both CRPS and NLL worsen, reflecting the gradual return to a genuinely difficult multimodal problem. Interval coverage remains reasonable across all noise levels, with 80% and 90% intervals close to their nominal coverage and 50% intervals slightly underdispersed.

This ablation shows that the BMDN responds in a principled way to side information about regimes. When a perfect disambiguating feature is available, compositional uncertainty nearly collapses and the model behaves like a standard unimodal regressor with small aleatoric noise. As the feature becomes unreliable, predictive performance degrades and the decomposition routes almost all additional variance into the compositional term, while keeping aleatoric and epistemic contributions relatively stable. This behavior matches the theoretical role of $C(x)$ as structural or branch–selection uncertainty.

5.5.4 A4: Effect of training set size on epistemic uncertainty. For a Bayesian model, epistemic uncertainty should shrink as the amount of data increases, while aleatoric and compositional uncertainty should be largely determined by the data generating process. In this ablation we vary the number of training samples for the BMDN and test whether the estimated epistemic component behaves consistently with this expectation.

We fix the BMDN architecture, priors, and optimization hyperparameters, and vary only the number of training points

$$n_{\text{train}} \in \{4{,}000, 6{,}000, 8{,}000, 12{,}000, 16{,}000, 20{,}000\}.$$

For each value of n_{train} we train a separate BMDN from scratch and evaluate on the same held–out validation set. From the predictive mixture we compute, for each validation input x, the aleatoric, compositional, and epistemic variances, and the total Monte Carlo predictive variance. We then report the mean of each quantity over the validation set, together with RMSE, NLL, CRPS, and R^2.

Figure 6 shows the mean epistemic variance as a function of training set size.

Standard metrics improve as training data increase and then saturate. RMSE decreases from about 0.191 at 4,000 samples to about 0.183 at 12,000 samples, while R^2 increases from roughly 0.46 to about 0.51 over the same range. CRPS and NLL follow the same pattern, with clear gains up to 8k–12k examples and small fluctuations thereafter.

The uncertainty decomposition shows the expected behavior. The mean epistemic variance decreases sharply with n_{train}: at 4,000 samples it is comparable in magnitude to the aleatoric term, while at 16,000 to 20,000 samples it is only a small fraction of the aleatoric variance and two orders of magnitude smaller than the compositional term. Aleatoric variance remains approximately constant across all training sizes, fluctuating in a narrow band that matches the injected noise level.

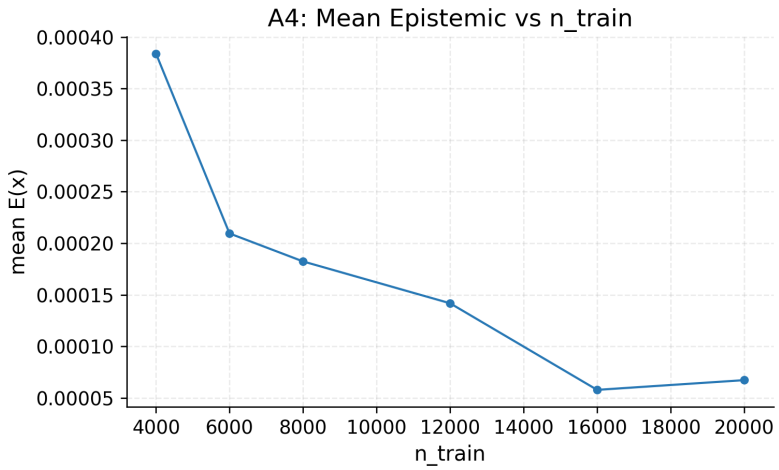


Figure 6. Ablation A4: mean epistemic variance $E(x)$ on the validation set as a function of the number of training samples n_{train}.

Compositional variance dominates the decomposition numerically for all n_{train}, with only a mild trend as the model refines its representation of the multimodal structure.

The mean Monte Carlo variance closely matches the sum of the three components at every sample size, which validates the numerical implementation of the decomposition. This ablation confirms that increasing the number of training samples has a strong and interpretable effect on the epistemic component, while leaving aleatoric and compositional uncertainty essentially determined by the underlying data generating process.

5.5.5 A5: Sensitivity to the Bayesian prior scale. The BMDN introduces a Bayesian last layer whose weight prior is controlled by a scale parameter c (denoted `prior_c` in the code). A very small c tightly regularizes the weights around zero, while a large c yields a diffuse prior and allows substantial posterior weight variability. Since this layer is the main source of epistemic uncertainty in the model, we test how sensitive both predictive performance and the $A/C/E$ decomposition are to the choice of c.

We keep the BMDN architecture, training schedule, and dataset fixed, and vary only the prior scale

$$c \in \{0.05, 0.1, 0.2, 0.5, 1, 2, 5, 10\}.$$

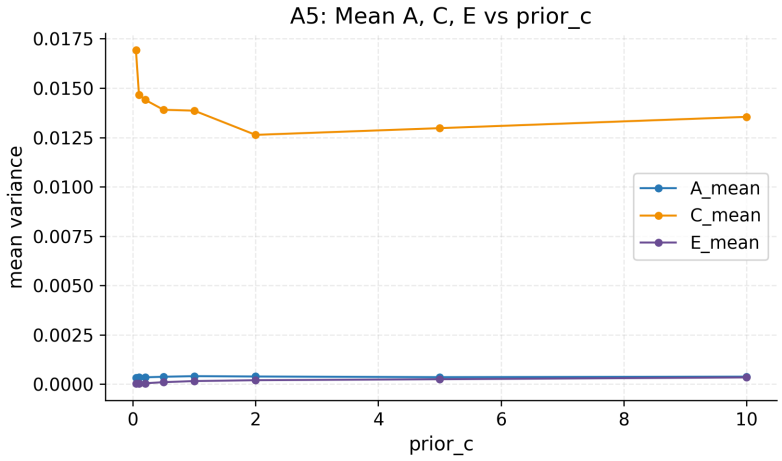


Figure 7. Ablation A5: mean aleatoric $A(x)$, compositional $C(x)$, and epistemic $E(x)$ variance as a function of the prior scale c for the Bayesian last layer.

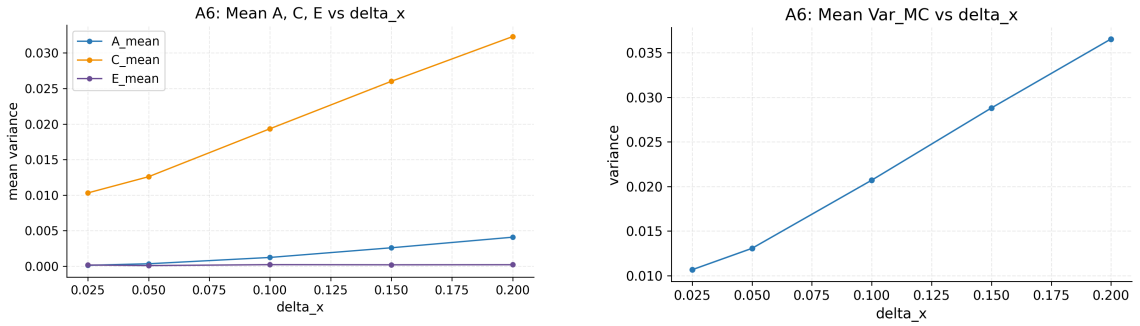
For each value of c we train the model once and evaluate on the validation set, reporting RMSE, CRPS, NLL, R^2, the Monte Carlo predictive variance, and the mean $A(x)$, $C(x)$, and $E(x)$.

Figure 7 shows the effect of the prior scale on the decomposition.

Across two orders of magnitude in prior scale, the point prediction metrics change only slightly. RMSE varies in a very narrow band, R^2 remains stable, and CRPS exhibits only minor fluctuations. NLL is somewhat more sensitive, with the best values near $c = 0.1$ and gradual degradation for very diffuse priors, but even there the differences are modest.

Compositional uncertainty dominates the variance budget for every prior setting, accounting for roughly 95–98% of the total predictive variance. Its mean follows a shallow U–shaped pattern as a function of c: it is largest at the tightest prior, decreases to a minimum around $c = 1$ to 2, and then increases slightly again for very diffuse priors. Epistemic variance increases monotonically with the prior scale, roughly tenfold between $c = 0.05$ and $c = 10$, as expected from a broader weight prior. Aleatoric variance remains nearly invariant to the prior, reflecting its interpretation as data noise rather than a prior–dependent artifact.

The average Monte Carlo variance shows the same U–shaped dependence as the compositional term and matches the sum $A + C + E$ closely at all prior scales. This ablation shows that the BMDN's predictive performance is largely insensitive to the exact prior scale within a broad range and that the prior primarily controls



(a) ACE vs δ_x

(b) $\mathrm{Var}_{\mathrm{MC}}$ vs δ_x

Figure 8. Ablation A6: effect of input noise scale δ_x on the three-way variance decomposition and the total predictive variance. Left: mean aleatoric $A(x)$, compositional $C(x)$, and epistemic $E(x)$ variance as a function of δ_x. Right: mean Monte Carlo predictive variance $\mathbb{E}[\mathrm{Var}_{\mathrm{MC}}(x)]$ as a function of δ_x.

the tradeoff between compositional and epistemic uncertainty, while leaving aleatoric uncertainty almost unchanged.

5.5.6 A6: Sensitivity of BMDN to data noise. The three–way decomposition is intended to separate uncertainty due to data noise from structural ambiguity and parameter uncertainty. If this is correct, then increasing the observational noise in the problem should primarily increase the aleatoric term $A(x)$, with a smaller effect on the compositional term $C(x)$ and little to no effect on the epistemic term $E(x)$. This ablation tests whether the BMDN respects that separation.

Starting from the default BMDN configuration, we vary only the scale of the noise used to generate the observed inputs x from the latent variable y. Specifically, we draw

$$x = g(y) + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \delta_x^2),$$

and run separate training runs for $\delta_x \in \{0.025, 0.05, 0.10, 0.15, 0.20\}$. The architecture, priors, optimization hyperparameters, and train–validation–test splits are held fixed.

For each trained model we evaluate the three–way variance decomposition, the total predictive variance, and the standard metrics on the validation set. Figure 8 shows how the mean uncertainties and total variance respond to the noise scale.

The aleatoric term is almost zero at the smallest noise level and increases monotonically and sharply with δ_x, growing by roughly a factor of thirty between $\delta_x = 0.025$ and $\delta_x = 0.20$. The compositional term is an order of magnitude larger

than $A(x)$ at all noise levels and also increases smoothly with δ_x, roughly tripling over the sweep. The epistemic term remains very small throughout and shows only minor, non–monotonic fluctuations, with a scale two orders of magnitude smaller than $C(x)$ and one order of magnitude smaller than $A(x)$ at high noise. The sum $A + C + E$ closely matches the mean Monte Carlo variance for all δ_x, and both increase nearly linearly with the noise scale.

Predictive performance degrades in the expected way as the input noise grows: RMSE increases, R^2 decreases, and both NLL and CRPS worsen monotonically. There are no abrupt jumps or instabilities in any of the curves, which suggests that the BMDN posterior remains in a consistent regime and adapts to the harder data rather than collapsing or overfitting.

This ablation provides a clean sanity check of the intended semantics of the three uncertainty terms. Increasing observational noise primarily increases the aleatoric term, while leaving epistemic uncertainty nearly unchanged. Compositional uncertainty remains the dominant contributor to the total variance, which is appropriate for this strongly multimodal problem, and its gradual growth with δ_x reflects the increased ambiguity about which branch generated each observation.

5.6 Summary of synthetic validation

This chapter used a controlled synthetic problem to validate that the proposed Bayesian Mixture Density Network behaves as required for the real carbon flux task. We evaluated calibration using proper scoring rules and diagnostics that are standard in probabilistic forecasting Bröcker and Smith (2007); Gneiting et al. (2007); Gneiting and Raftery (2007); Kuleshov et al. (2018). The main conclusions are as follows.

First, the BMDN matches the point prediction performance of a heteroscedastic Gaussian regressor and a deterministic MDN, while retaining the probabilistic improvements of the mixture model. In a regime with ample data and a smooth target, there is little difference between these models in terms of RMSE and R^2, but the MDN and BMDN substantially improve probabilistic scores relative to a single Gaussian.

Second, in a setting where oracle uncertainties are available, the BMDN recovers the qualitative shape and quantitative scale of the aleatoric, compositional, and epistemic components. Compositional uncertainty peaks exactly where the inverse is multi–valued; aleatoric uncertainty tracks the injected input noise; and epistemic uncertainty is largest where data are structurally ambiguous and the model has

less effective support. The decomposition sums back to the Monte Carlo predictive variance, confirming numerical consistency.

Third, the BMDN achieves predictive distributions that are well calibrated according to reliability curves and PIT histograms, with calibration comparable to or slightly better than the best deterministic model. This is important if the uncertainty estimates are to be trusted in downstream applications.

Finally, a suite of ablations shows that the three components $A(x)$, $C(x)$, and $E(x)$ respond in ways that match their intended semantics under a wide range of controlled changes: mixture capacity, Bayesian depth, presence or absence of a disambiguating feature, training set size, prior scale, and data noise. Aleatoric uncertainty responds primarily to observational noise, compositional uncertainty to structural ambiguity and regime information, and epistemic uncertainty to data abundance and Bayesian modeling choices.

These synthetic results do not simply show that the BMDN "works" on a toy problem. They demonstrate that the three–way decomposition is identifiable, numerically stable, and interpretable in a setting that closely mirrors the structural difficulty of carbon flux prediction while providing ground truth for comparison. This justifies using the BMDN and its $A/C/E$ decomposition as a lens through which to analyze uncertainty in the real carbon flux experiments in Chapter VI.

CHAPTER VI
CARBON FLUX EXPERIMENTS ON REAL AMERIFLUX DATA

Figure 9. Eddy-covariance tower located in a coastal Oregon wetland.

This chapter uses material from: Anish Dulal and Jake Searcy. "Uncertainty-Aware Carbon Flux Estimation from Multispectral Landsat Imagery Using Mixture Density Networks," ICLR 2025 Workshop on Tackling Climate Change with Machine Learning *Dulal and Searcy (2025). Dulal led the development of the methodology, implemented all experiments, and wrote the manuscript. Searcy contributed to the conceptual framing, advised the experiments, and assisted with manuscript reviewing and editing.*

In the previous chapter we showed, on a fully controlled synthetic problem, that a Bayesian Mixture Density Network (BMDN) can approximately recover the desired three-way decomposition of predictive uncertainty into aleatoric, compositional, and epistemic components. The central question for this chapter is simple but critical:

> Given a realistic, messy, multimodal carbon flux dataset, does the same BMDN architecture still produce a three-way decomposition that is consistent, interpretable, and scientifically useful?

To answer this, we train a BMDN on AmeriFlux tower data collocated with Landsat and meteorological covariates, and then analyze its behavior under three regimes: in-distribution validation, temporal extrapolation, and spatial or domain shift. We evaluate standard predictive metrics and decomposed uncertainty, examine

variation across land-cover classes, and assess calibration. We also study feature-level attributions for the mean and each uncertainty component, and explore both hyperparameter sensitivity and structural ablations.

Throughout this chapter the model is exactly the one defined in the methods chapter. It is a deterministic feature extractor with several hidden layers, followed by a single Bayesian hidden layer, and then a deterministic mixture density head. Only that last hidden layer is treated in a Bayesian way. The MDN head is always deterministic.

6.1 Data and evaluation protocol

We use the AMERI-FAR25 dataset Searcy et al. (2025), which combines 439 site-years of AmeriFlux tower data with Landsat scenes. This dataset consists of collocated AmeriFlux tower NEE measurements and remote sensing plus meteorological covariates Chu et al. (2021); Delwiche et al. (2024); Papale (2020); Pastorello et al. (2020) and the same tower network as in our earlier workshop paper Dulal and Searcy (2025). The target variable is net ecosystem exchange (NEE) in the $\mu\text{mol} \cdot \text{m}^{-2} \cdot \text{s}^{-1}$ units, with the sign convention that negative values denote net carbon uptake.

The final dataset comprises 439 site-years of eddy-covariance (EC) tower data from 209 AmeriFlux sites. Starting from the AmeriFlux BASE product, sites with missing measurements or incomplete metadata were removed, and the remaining time series were cleaned and harmonised using the ONEFlux pipeline. On the remote-sensing side we use Landsat 8 and 9 Level-1 scenes. These provide eleven spectral bands, including nine surface-reflectance bands at 30 m and two thermal infrared bands originally at 100 m resolution but resampled to 30 m. Across all sites and years, 45,124 Landsat scenes contribute at least one valid observation in the vicinity of a tower. Figure 9 shows a representative eddy-covariance tower in a coastal Oregon wetland.

The covariate vector combines the following groups. For each scene we extract a 5×5 pixel patch centered on the tower location and process it to account for clouds and missing data. All bands except the panchromatic band, which provides broadband measurements at higher spatial resolution, are retained. Each patch is summarised into a feature vector by concatenating the spectral bands with four solar and viewing geometry angles (solar/sensor, azimuth/zenith), along with meteorological and ancillary drivers described below.

Landsat surface reflectance and thermal bands

- Band 1 Coastal aerosol
- Band 2 Blue
- Band 3 Green
- Band 4 Red
- Band 5 Near Infrared (NIR)
- Band 6 Shortwave Infrared (SWIR) 1
- Band 7 Shortwave Infrared (SWIR) 2
- Band 9 Cirrus
- Band 10 Thermal Infrared (TIRS) 1
- Band 11 Thermal Infrared (TIRS) 2

Solar and viewing geometry

- SZA (solar zenith angle)
- SAA (solar azimuth angle)
- VZA (view zenith angle)
- VAA (view azimuth angle)

Meteorological and surface variables

- SW_IN incoming shortwave radiation
- TA air temperature
- RH relative humidity
- USTAR friction velocity
- Wind speed
- Wind direction
- HEIGHT (tower height)
- TIME_MASK, which encodes whether current value is replace by past values or not.

All continuous inputs and the NEE target are normalized using a linear transformation based on training statistics only, following the same conventions as in the synthetic chapter.

6.1.1 Data splits and evaluation regimes. We use three conceptually distinct evaluation regimes that correspond to different kinds of distribution shift, following common practice in flux up-scaling and remote-sensing validation Jung et al. (2020); Pastorello et al. (2020).

Train and validation: in distribution Training and validation sets contain data from the same sites and the same range of years and are used for fitting model weights and tuning hyperparameters. After constructing the full dataset, we first set aside test splits (described below) and then draw a random 20% of the remaining points into a validation set (`val`), keeping the rest for training. This approximates the standard i.i.d. regime used for model selection.

Future years at known sites: temporal shift To probe robustness under temporal drift, the `test_future` split contains the last year of data from every site with multiple years of observations. These held-out years are removed entirely from the training and validation sets. The sites themselves still appear in training, so `test_future` measures extrapolation in time rather than in space.

New sites: spatial shift To assess spatial generalisation, the `test_site` split contains whole sites that never appear in train or validation. Sites are grouped by their IGBP land-cover code. For each IGBP class with at least ten sites, we randomly assign 20% of sites to `test_site` and retain the remaining 80% for train/validation. This preserves coverage across ecosystem types while ensuring that `test_site` consists entirely of unseen locations.

All hyperparameter choices and model selection decisions are based solely on the validation split. The `test_future` and `test_site` splits are reserved for reporting and analysis.

6.2 Overall performance and uncertainty decomposition

Table 2 summarizes the performance of the final tuned BMDN on the four main splits. We report root mean squared error (RMSE), coefficient of determination (R^2), negative log likelihood (NLL), continuous ranked probability score (CRPS), empirical coverage of central 50, 80, and 90 percent predictive intervals, and the mean aleatoric, compositional, and epistemic standard deviation components.

Several patterns are worth highlighting.

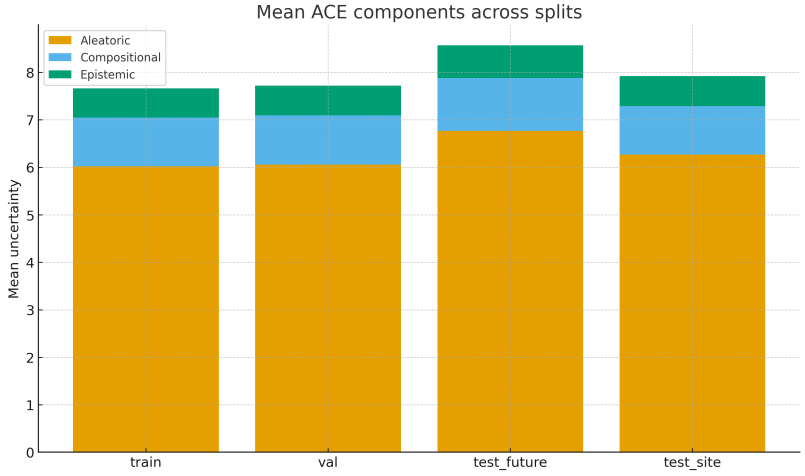
First, predictive accuracy remains strong across all regimes. On the validation split the model explains about 67% of the variance with an RMSE of approximately 2.76. Under temporal shift in `test_future`, R^2 is close to 0.70 and RMSE increases modestly. Under spatial shift in `test_site`, R^2 drops to about 0.63 with RMSE around 3.1. The degradation is noticeable but controlled, which is encouraging given the complexity of the task.

Table 2. Overall performance and uncertainty decomposition of the final BMDN across training, validation, and test splits. Cov 50, Cov 80, and Cov 90 denote empirical coverage of central predictive intervals at the corresponding nominal levels. A_sd, C_sd, and E_sd denote the mean aleatoric, compositional, and epistemic standard deviations of the predictive distribution, averaged over samples in each split.

Split	RMSE	R^2	NLL	CRPS	Cov 50	Cov 80	Cov 90	A_sd	C_sd	E_sd
train	2.5271	0.7320	1.7059	1.0571	0.5887	0.8631	0.9321	2.0262	1.0153	0.6854
val	2.7605	0.6745	1.7856	1.1570	0.5514	0.8394	0.9183	2.0422	1.0277	0.6946
test_future	2.9021	0.7002	1.9471	1.2957	0.5133	0.8069	0.8970	2.1647	1.0809	0.7330
test_site	3.1233	0.6286	1.8434	1.3071	0.4857	0.7905	0.8883	2.0273	1.0264	0.6878

Second, the ACE decomposition behaves in a manner that mirrors the synthetic chapter. Throughout all splits, the aleatoric variance A is the dominant contribution, as expected for tower-scale NEE where measurement noise, footprint mismatch, and unresolved subgrid processes are substantial. The compositional variance C is smaller but clearly nonzero and is largest under `test_future`. The epistemic variance E is the smallest component, but it increases from train and validation to `test_future` and remains at a similar level magnitude to train and validation on `test_site`.

Third, calibration is broadly satisfactory. Central 80% prediction intervals cover roughly 0.79–0.86 of observations across splits, and 90% intervals cover roughly 0.89–0.93. Coverage is close to nominal, with some splits slightly over-covering and others slightly under-covering, rather than showing strong overconfidence.



6.3 Variation across land cover classes

The global averages in Table 2 hide large and scientifically meaningful differences across land cover types. In this section we group results by the IGBP class associated with each site and examine how performance and the ACE components change.

6.3.1 Temporal shift: future years at known sites. Table 3 reports per class metrics for the `test_future` split.

Table 3. Per IGBP performance and uncertainty components on the `test_future` split. A_sd, C_sd, and E_sd denote the mean aleatoric, compositional, and epistemic standard deviations of the predictive distribution within each IGBP class.

IGBP	RMSE	R^2	count	A_sd	C_sd	E_sd
CRO	3.26	0.6476	48,883	2.12	1.22	0.73
CSH	1.98	0.7954	37,624	1.75	0.83	0.57
DBF	3.44	0.7469	30,403	2.61	1.30	0.91
DNF	4.42	0.2181	5,912	2.80	1.65	0.99
EBF	5.41	-0.2556	3,074	3.59	1.99	1.42
ENF	2.83	0.7342	68,976	2.43	1.22	0.80
GRA	3.26	0.5899	68,627	2.23	1.11	0.77
MF	3.57	0.7930	13,159	3.21	1.64	1.13
OSH	1.38	0.5310	31,011	1.33	0.55	0.44
SAV	1.00	0.5904	12,626	1.22	0.51	0.39
SNO	2.12	0.2784	5,186	1.25	0.48	0.38
WAT	4.87	-0.0234	2,696	2.85	1.50	0.92
WET	2.94	0.7526	71,493	2.33	1.13	0.79
WSA	1.19	0.4113	13,546	1.32	0.60	0.45

For well instrumented forest and wetland classes (DBF, ENF, WET, MF) predictive performance remains strong and R^2 stays high under future years, but both A and E increase relative to validation. Epistemic variance is particularly large in classes with relatively few sites and complex structure (DNF, EBF, MF, WAT), which is exactly where the model is expected to be least certain about future conditions. Open shrublands and savanna classes (OSH, SAV, WSA) again show clearly smaller absolute uncertainty and lower E, while grasslands (GRA) have intermediate uncertainties that are still lower than in structurally more complex forests.".

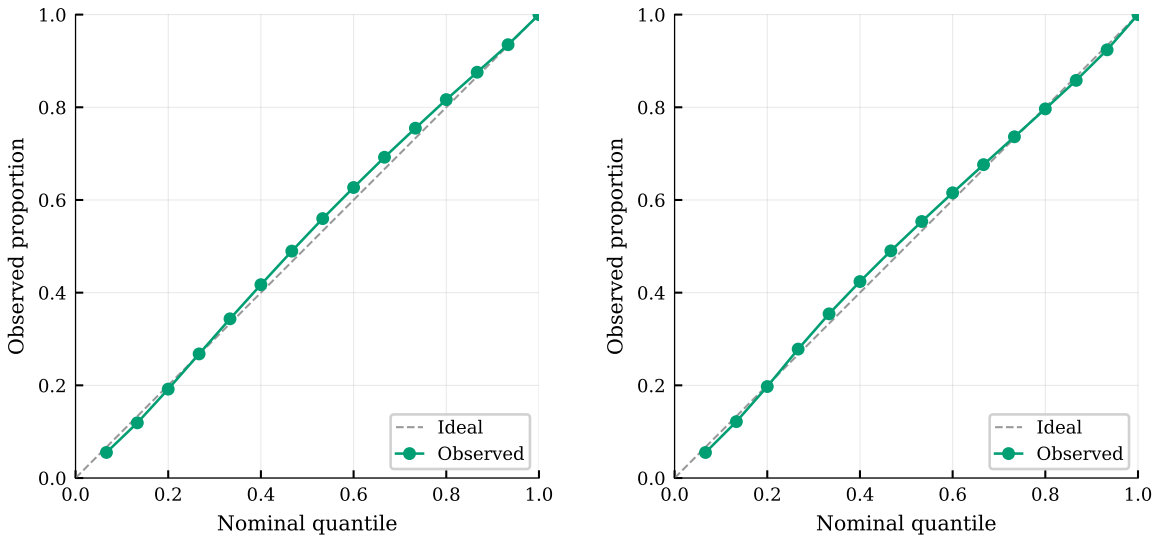
6.3.2 Spatial shift: new sites. On `test_site` we only have IGBP classes corresponding to sites that were held out in space. Table 4 shows the per class metrics.

Table 4. Per IGBP performance and uncertainty components on the `test_site` split. A_sd, C_sd, and E_sd denote the mean aleatoric, compositional, and epistemic standard deviations of the predictive distribution within each IGBP class.

IGBP	RMSE	R^2	count	A_sd	C_sd	E_sd
CRO	5.02	0.4827	225,916	2.85	1.78	1.04
DBF	3.95	0.7219	176,235	2.82	1.42	0.96
ENF	2.84	0.6102	218,025	2.21	1.07	0.74
GRA	2.38	0.5955	310,081	1.81	0.90	0.60
OSH	1.15	0.2704	456,222	1.16	0.49	0.37
WET	3.75	0.6605	227,503	2.46	1.18	0.83

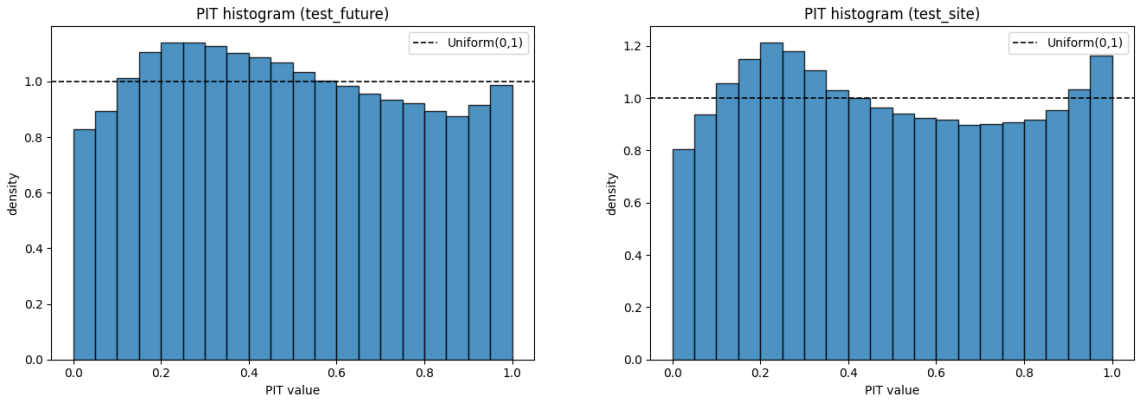
Spatial or domain shift affects both accuracy and uncertainty. For CRO, DBF, and WET, RMSE increases and A grows substantially relative to validation. This reflects new local conditions and site level heterogeneity. Epistemic variance E is also higher for these classes than on train and validation, since the model faces ecosystems and tower configurations it has not seen before. In contrast, GRA remains relatively easy, with small A and E and R^2 above 0.45 even for unseen sites.

6.4 Calibration of predictive distributions



(a) `test_future` (b) `test_site`

Figure 10. Reliability curves of the final BMDN on the `test_future` and `test_site` splits. Each panel shows empirical coverage as a function of nominal coverage for fifteen quantiles, with the $y = x$ line as reference.



(a) `test_future` (b) `test_site`

Figure 11. Probability integral transform (PIT) histograms for the `test_future` and `test_site` splits.

Calibration is central for any uncertainty quantification method Gneiting et al. (2007); Gneiting and Raftery (2007). For each test split we compute reliability curves by comparing nominal quantiles with empirical coverage, as shown in Figure 10. We use fifteen equally spaced nominal levels between 0.0667 and 1.0, following standard practice in probabilistic regression Bröcker and Smith (2007); Kuleshov et al. (2018).

On test_future, empirical coverage tracks the nominal levels closely across the entire range. For example, at a nominal level of 0.80 the observed coverage is about 0.817, and at nominal 0.8667 and 0.9333 the observed coverages are about 0.876 and 0.935, respectively. This corresponds to very mild overcoverage at intermediate levels, which is consistent with the integrated 80 percent and 90 percent central interval coverages of 0.807 and 0.897.

On test_site, the reliability curve is again close to the diagonal. At intermediate nominal levels it tends to lie slightly above the diagonal, while in the upper tail it lies slightly below it. For instance, the observed coverages at nominal 0.80, 0.8667, and 0.9333 are approximately 0.797, 0.858, and 0.924. The central 80 percent and 90 percent intervals cover about 0.791 and 0.888 of observations, respectively. Overall, the model is slightly underconfident on some ranges and slightly overconfident on others, but deviations from perfect calibration are modest in both test regimes.

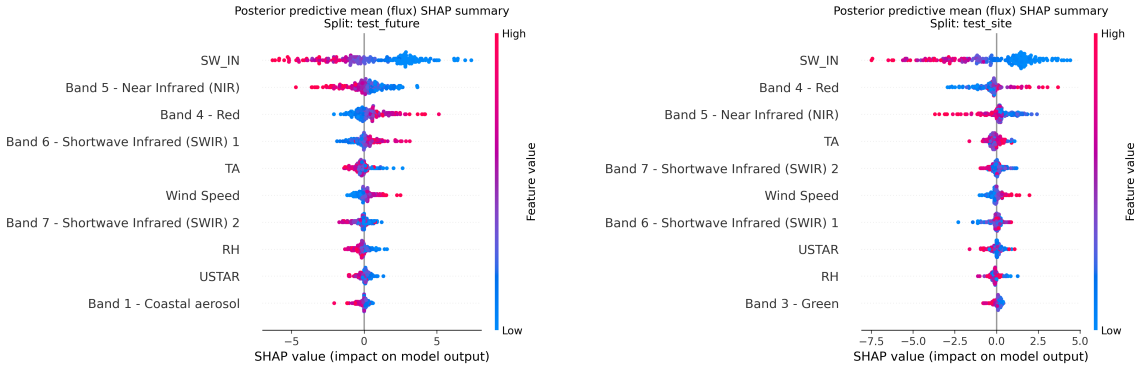
We also compute probability integral transform (PIT) histograms by evaluating the posterior predictive CDF at each observed NEE value and histogramming the resulting PIT values. Figure 11 shows these histograms for the two test splits. They

provide a complementary diagnostic of calibration and can reveal underdispersion, overdispersion, or other systematic biases Gneiting et al. (2007); Kuleshov et al. (2018).

6.5 Feature-level explanations via SHAP

To make the BMDN behavior interpretable, we compute SHAP values for the predictive mean and for each of the aleatoric, compositional, and epistemic variance components on the two test splits Covert, Lundberg, and Lee (2021); Lundberg and Lee (2017); Shapley (1953). We use a sampling-based SHAP approximation applied to the trained network, with background points drawn from the training distribution. This yields feature importance rankings that can be compared across the four quantities.

SHAP values should be interpreted with some caution in this setting. Many predictors (for example, spectral bands and radiation variables) are correlated, and SHAP inherently attributes marginal contributions under a particular conditional independence model. As a result, absolute magnitudes of SHAP scores are less meaningful than relative rankings within a panel or qualitative shifts across splits and uncertainty components. We therefore focus on robust patterns—which variables consistently appear among the top contributors for the mean, A, C, or E, rather than on fine differences in score.



(a) Predicted flux, `test_future`

(b) Predicted flux, `test_site`

Figure 12. SHAP summary plots for the predictive mean flux on the `test_future` and `test_site` splits. These panels show which input features most strongly influence the point predictions.

6.5.1 Predictive mean. For both test splits the mean flux prediction is controlled mainly by SW_IN and the NIR and red bands, with SWIR bands and

air temperature (TA) providing secondary contributions (Figure 12). SW_IN has by far the widest SHAP range, indicating that illumination is the dominant driver of the posterior mean. The spectral bands show clear color gradients, consistent with the model using reflectance as a proxy for canopy state and soil exposure. TA has a moderate but coherent effect, while RH, wind speed, USTAR and the remaining bands appear with smaller spreads and lower ranks. Overall, the BMDN bases its central prediction on physically plausible controls that are stable across both the temporal and spatial test splits.

6.5.2 Aleatoric variance. Aleatoric uncertainty A is also most sensitive to SW_IN and the red and NIR bands (top three variables in both splits), with RH and the SWIR bands next in importance. TA, USTAR and wind speed contribute, but they sit in the lower half of the ranking and have visibly smaller SHAP spreads. The point clouds for SW_IN and the spectral bands widen at extreme feature values, which indicates that observation noise and sub-footprint variability are larger under particular combinations of illumination and canopy state. Thus A is structured rather than uniform, yet it is not primarily controlled by the turbulence metrics.

6.5.3 Compositional variance. Compositional uncertainty C is even more tightly linked to the spectral information. On `test_future`, the three SWIR and NIR bands occupy the top positions, with SWIR2 leading, while on `test_site` the strongest drivers are red, NIR and SWIR2. SW_IN and meteorological variables only appear in the middle or lower part of the ranking and have narrower SHAP distributions. The color gradients for the spectral bands show that particular reflectance regimes produce large positive or negative contributions to C, whereas other regimes leave it close to its baseline. This pattern is consistent with C capturing heterogeneity in surface composition that is expressed in the optical bands and only weakly modulated by the local meteorology.

6.5.4 Epistemic variance. Epistemic uncertainty E depends on almost the same set of drivers as the mean, but with much smaller SHAP magnitudes. SW_IN together with the red, NIR and SWIR2 bands are the leading predictors on both splits, followed by RH and TA; USTAR, wind speed and the remaining bands are less influential. The dependence on SW_IN and reflectance is again monotonic in color, and the point clouds broaden mainly at the extremes of illumination and canopy state. This indicates that the model is most uncertain in precisely those parts of feature space where the physically important drivers also vary strongly, rather than

in arbitrary directions. Geometry variables do not appear among the top ten inputs, so the dominant structure in E is controlled by radiation and surface properties rather than by viewing or solar angle.

6.6 Hyperparameter sweeps and robustness

Before locking in a final BMDN configuration, we perform systematic sweeps over key hyperparameters. All sweeps evaluate solely on the validation splits. The goal is twofold. First, we seek a region in hyperparameter space with strong predictive performance and good calibration. Second, we check that the qualitative ACE behavior is robust rather than being an artifact of a single parameter choice.

In all sweeps we track RMSE, R^2, NLL, CRPS, and the mean A, C, and E components on the validation data.

6.6.1 Number of mixture components. The number of mixture components K controls how many distinct modes the MDN can represent. The sweep ranges from $K = 1$ to $K = 10$. The detailed table is given in the appendix; here we summarize the main trends, which are visualized in Figure 14.

When $K = 1$ the model collapses to a single Gaussian per input. Validation R^2 is about 0.59 and NLL is worst among all settings. Compositional variance is zero by construction, so all residual uncertainty is forced into A and E.

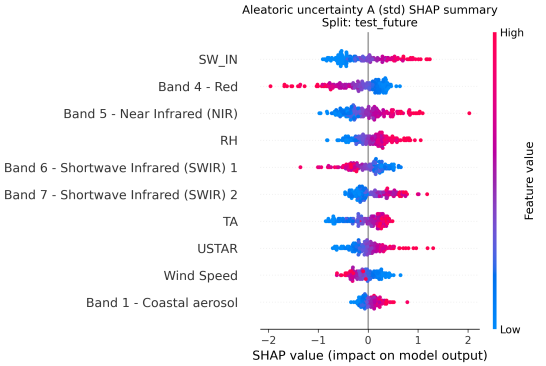
As K increases from 2 to about 5 there is a clear improvement. RMSE and NLL decrease, R^2 increases to about 0.62, and C becomes nonzero and grows steadily, while A decreases. This is exactly what we expect when the model starts to represent multimodality explicitly.

For K between 5 and 9 the gains in RMSE and NLL become marginal. Validation R^2 improves only slightly. Compositional variance continues to grow, while A shrinks and E remains relatively stable. Very large K moves more variance into C without substantial benefits in standard metrics, which complicates interpretation and increases computational cost.

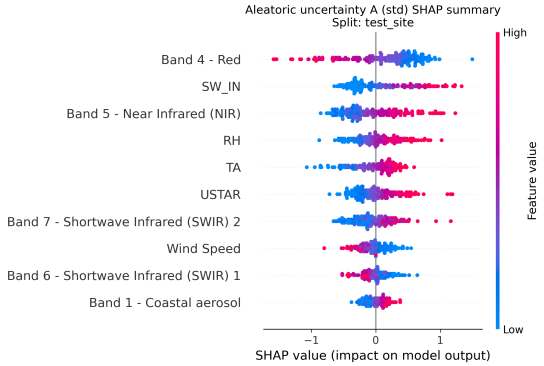
The final model is chosen from this plateau of moderate $K = 7$ where predictive performance is near optimal and the ACE decomposition remains interpretable.

6.6.2 Prior scale. The prior scale on the Bayesian hidden layer controls how strongly weights are regularized. We sweep scales from 1.0 (strong prior) to 50.0 (weak prior). Representative results on validation are shown in Figure 15.

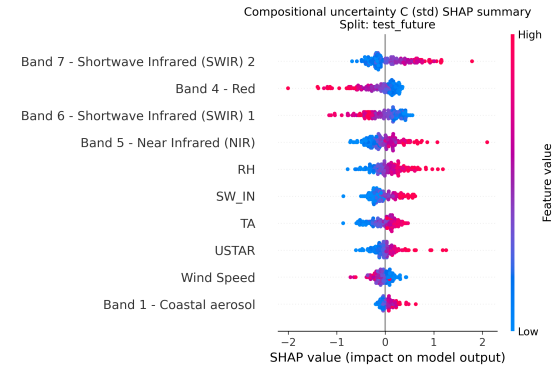
Very strong priors underfit and give worse R^2, RMSE, and NLL. For intermediate scales between about 15 and 30, validation metrics are improved and A and C take



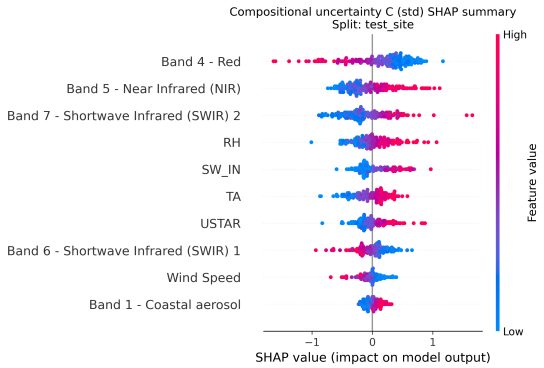
(a) Aleatoric, `test_future`



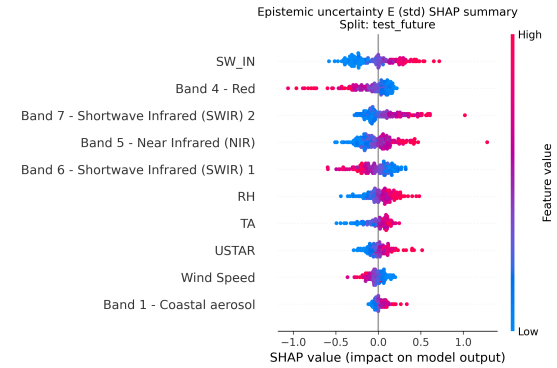
(b) Aleatoric, `test_site`



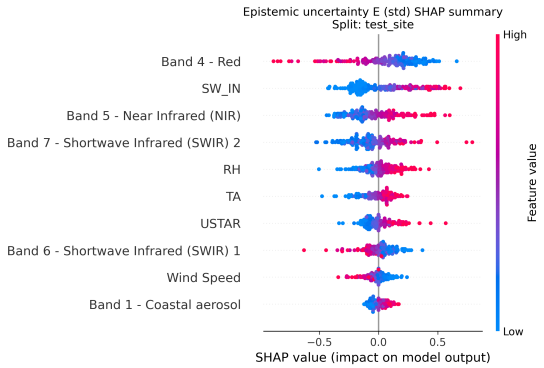
(c) Compositional, `test_future`



(d) Compositional, `test_site`



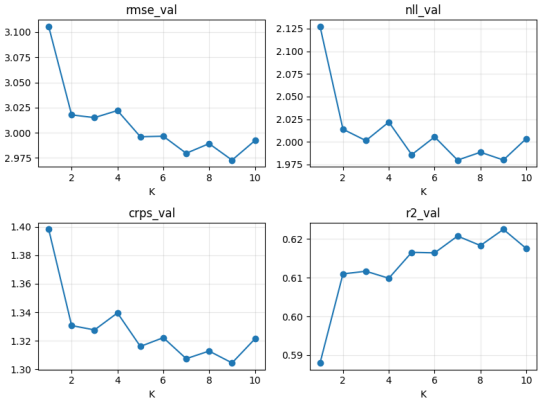
(e) Epistemic, `test_future`



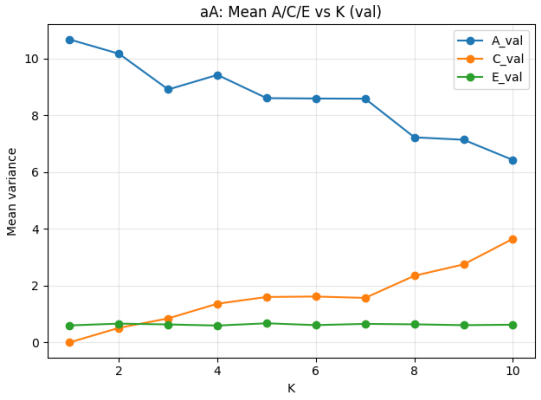
(f) Epistemic, `test_site`

Figure 13. SHAP summary plots for the `test_future` and `test_site` splits, decomposed by uncertainty type. Rows correspond to aleatoric, compositional, and epistemic predictive standard deviation, while columns compare the two test splits. Features are ordered by average absolute SHAP value within each panel.

reasonable values. Extremely weak priors do not provide further gains and can slightly

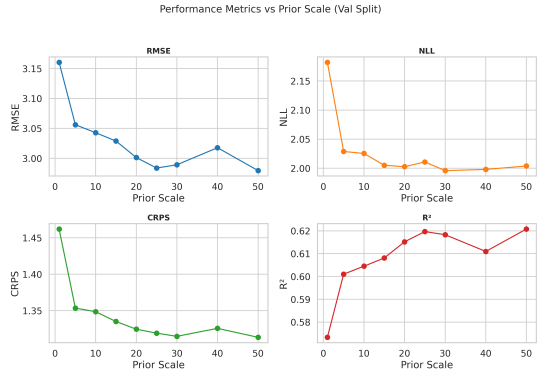


(a) Validation performance metrics

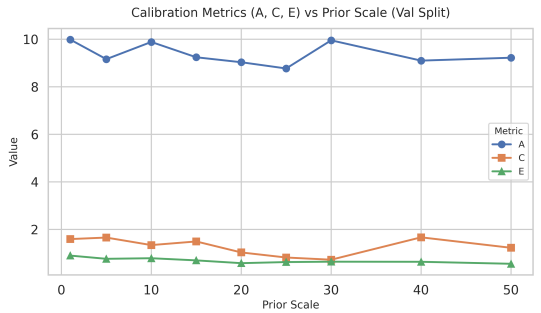


(b) Mean uncertainties on validation

Figure 14. Effect of the number of mixture components K on the BMDN using the validation split. Left: RMSE, NLL, CRPS, and R^2 used to tune K. Right: mean aleatoric, compositional, and epistemic variance averaged over the validation data.



(a) Validation performance metrics



(b) Mean uncertainties on validation

Figure 15. Effect of the Bayesian hidden-layer prior scale on the BMDN using the validation split. Left: RMSE, NLL, CRPS, and R^2 as functions of the prior scale. Right: mean aleatoric, compositional, and epistemic variance averaged over the validation data.

harm the decomposition. There is a broad plateau of good performance for moderate prior strengths, and the final model uses a value of 30.0.

6.6.3 Sigma regularization. The parameter `sigma_reg_scale` controls explicit regularization on the mixture component variances. The sweep indicates that too little regularization allows variances to grow excessively, which leads to worse NLL and a tendency toward overdispersion. Too strong regularization forces variances to be too small, which harms calibration and inflates E when the model cannot fit data

through variance. A mild positive regularization around `sigma_reg_scale` between about 0.00085 and 0.0014 yields a good trade-off between NLL, CRPS, and coverage. The final model uses a value of 0.00085.

6.6.4 Max temperature. The `max_temperature` hyperparameter sets a cap for variance scaling before a penalty kicks in. We sweep values between 2.0 and 10.0. Very low caps such as 2.0 restrict variance adaptation and give slightly worse NLL and higher CRPS. Very high caps do not significantly improve NLL and can make calibration a little less stable. Intermediate caps around 4.0 to 8.0 provide the best combination of NLL, CRPS, and near-nominal coverage. The final configuration uses a value of 4.0.

6.6.5 Pi entropy penalty. The `pi_entropy_scale` penalizes overly concentrated or degenerate mixture weight distributions and helps avoid collapsed components. The sweep ranges from small values (for example 0.0007) to larger regularization.

Very small penalties allow highly concentrated mixture weights. This can sometimes reduce NLL but tends to produce large C and can slightly reduce robustness of calibration. Intermediate penalties around 0.0010 to 0.0015 yield a good balance. Mixture weights are neither fully collapsed nor completely flat. NLL and CRPS are competitive and coverage stays near nominal. Very large penalties would overly discourage confident mixture assignments and inflate A.

The final model uses a pi entropy of 0.0014.

6.7 Structural ablations on features and Bayesian depth

The sweeps above change scalar hyperparameters while keeping the overall structure fixed. In this section we consider structural ablations that change either the feature set or the depth of the Bayesian part of the network. These are key for understanding what each architectural component contributes to the behavior of the ACE decomposition.

6.7.1 Feature family ablations. We compare five configurations.

- **full**: all spectral, meteorological, and geometry features
- **no_spectra**: remove all spectral bands, keep meteorology and geometry
- **no_meteo**: remove meteorology, keep spectral and geometry
- **spectra_only**: spectral bands only
- **meteo_only**: meteorology and geometry only

Table 5 summarizes the results on both test splits.

Table 5. Effect of removing or isolating feature families on predictive performance and ACE decomposition for `test_future` and `test_site`. A_sd, C_sd, and E_sd denote the mean aleatoric, compositional, and epistemic standard deviations of the predictive distribution.

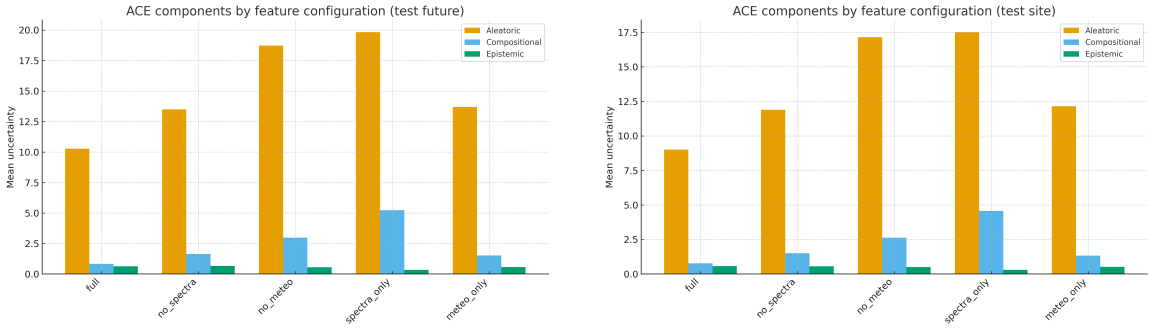
Mode	Split	RMSE	R^2	NLL	CRPS	A_sd	C_sd	E_sd
full	test_future	3.06	0.6657	2.1055	1.4165	3.20	0.91	0.79
full	test_site	3.06	0.6440	1.9181	1.2972	3.00	0.88	0.75
no_spectra	test_future	3.64	0.5290	2.2336	1.6763	3.67	1.28	0.81
no_spectra	test_site	3.49	0.5362	2.0211	1.4957	3.45	1.23	0.75
no_meteo	test_future	4.57	0.2552	2.4607	2.1294	4.33	1.73	0.74
no_meteo	test_site	4.34	0.2814	2.2282	1.8648	4.14	1.62	0.71
spectra_only	test_future	4.85	0.1644	2.5269	2.2673	4.45	2.29	0.58
spectra_only	test_site	4.62	0.1879	2.2938	1.9918	4.19	2.14	0.55
meteo_only	test_future	3.63	0.5298	2.2380	1.6756	3.70	1.23	0.76
meteo_only	test_site	3.49	0.5357	2.0378	1.4971	3.49	1.16	0.72

The full model is best in both R^2 and probabilistic metrics, which confirms that both spectral and meteorological features are necessary to capture flux dynamics. Removing spectral features (no_spectra, meteo_only) hurts R^2 and NLL and inflates both A and C. Without spectral information the model cannot distinguish different vegetation states, so both noise and multimodality increase. Removing meteorological features (no_meteo, spectra_only) is even more damaging. R^2 collapses toward 0.25 and A and C become extremely large. Spectra and geometry alone cannot capture short-term flux variability.

Compositional variance C is maximal in the spectra_only setting, which makes sense. Many distinct flux regimes share similar reflectance patterns when meteorology and management are not conditioned on. Epistemic variance remains relatively small in all settings. Most of the uncertainty is explained as A and C even when features are missing. This is consistent with E reflecting weight posterior uncertainty rather than being a generic residual. Figure 16 visualizes how the ACE components respond to these feature-family ablations on the two test splits.

6.7.2 Bayesian depth. The second structural ablation varies how many hidden layers are treated in a Bayesian way. The network has a deterministic feature extractor of several layers followed by a hidden layer directly feeding the MDN head. In all cases the MDN head itself is deterministic.

We compare three configurations.



(a) Temporal shift (test future)

(b) Spatial shift (test site)

Figure 16. Effect of removing meteorological and spectral feature groups on mean aleatoric (A), compositional (C), and epistemic (E) variances for test future and test site splits. Removing spectra or meteorology substantially increases C and A while leaving E relatively small, which is consistent with the interpretation of C as capturing unresolved mixture structure induced by missing covariates.

- **last**: only the last hidden layer before the MDN is Bayesian
- **last_two**: the last two hidden layers are Bayesian
- **last_three**: the last three hidden layers are Bayesian

Table 6 reports the results.

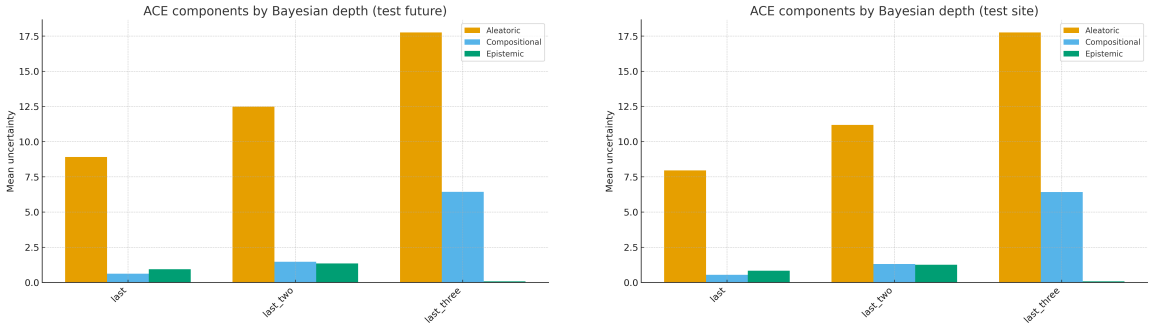
Table 6. Effect of Bayesian depth on predictive performance and ACE decomposition. In all cases the MDN head is deterministic. A_sd, C_sd, and E_sd denote the mean aleatoric, compositional, and epistemic standard deviations of the predictive distribution.

Mode	Split	RMSE	R^2	NLL	CRPS	A_sd	C_sd	E_sd
last	test_future	3.15	0.6468	2.1929	1.4820	3.29	1.19	0.83
last	test_site	3.11	0.6322	2.0375	1.3573	3.13	1.13	0.80
last_two	test_future	3.46	0.5733	2.3136	1.6541	3.50	1.50	1.03
last_two	test_site	3.34	0.5762	2.1359	1.4884	3.36	1.43	0.99
last_three	test_future	5.32	−0.0072	2.7591	2.6476	4.18	2.56	0.35
last_three	test_site	5.14	−0.0073	2.5404	2.3875	4.19	2.56	0.34

A single Bayesian last layer (last) delivers the best predictive performance and the most reasonable ACE decomposition. Aleatoric variance is dominant, compositional variance is moderate, and epistemic variance is small but nonzero. With two Bayesian layers (last_two), all metrics degrade and all three variance components inflate. The model becomes harder to train and the variational approximation must capture

uncertainty in a much larger weight space. With three Bayesian layers (last_three) the model essentially breaks. R^2 turns negative, RMSE explodes, and C grows to very large values while E nearly collapses.

For this dataset and architecture, a Bayesian last hidden layer is therefore the sweet spot. It injects enough epistemic flexibility to capture meaningful uncertainty about the mapping from features to mixture parameters, while avoiding the training instability and degeneracy that arise when too many layers are Bayesian. Figure 17 summarizes how the ACE components change with Bayesian depth under temporal and spatial shift.



(a) Temporal shift (test future)

(b) Spatial shift (test site)

Figure 17. Impact of Bayesian depth on mean aleatoric (A), compositional (C), and epistemic (E) variances. The configuration with a single Bayesian hidden layer (last only) offers a good trade-off between calibration and stability, while making more layers Bayesian inflates A and C and eventually destabilizes training.

6.8 Summary of carbon flux experiments

The carbon flux experiments in this chapter extend the ACE decomposition from a controlled synthetic problem to a multi-site AmeriFlux NEE dataset with 209 towers, 439 site-years, and thousands of Landsat-constrained samples spanning diverse IGBP land-cover classes. Several conclusions emerge.

First, the BMDN with a single Bayesian hidden layer and a moderate number of mixture components achieves competitive predictive performance on a challenging real-world task. Validation R^2 is about 0.67 with RMSE around 2.76. Under temporal shift (`test_future`), R^2 increases slightly to about 0.70 with a modest rise in RMSE, while under spatial shift to unseen sites (`test_site`), R^2 drops to roughly 0.63 and RMSE increases to about 3.1. The degradation under spatial shift is noticeable but controlled given the heterogeneity of the tower network.

Second, the qualitative behavior of the ACE components carries over from the synthetic chapter. On average across splits and land-cover types, aleatoric variance is dominant. Compositional variance is moderate on average and largest in structurally complex and multimodal ecosystems such as croplands, mixed forests, and wetlands. Epistemic variance is smallest but increases in underrepresented regimes and under both temporal and spatial shift. These patterns are qualitatively consistent with the theoretical decomposition and with the synthetic experiments.

Third, calibration is good. Reliability curves lie slightly above the diagonal, and central 80 percent prediction intervals cover roughly 0.79–0.86 of observations across splits, and 90 percent intervals cover roughly 0.89–0.93. Coverage is close to nominal, with small deviations in both directions, rather than showing strong overconfidence or severe undercoverage.

Fourth, feature-level SHAP analyses show that the predictive mean and all three uncertainty components are driven mainly by a physically interpretable set of variables, in particular SW_IN and the visible, NIR, and SWIR bands, with air temperature, humidity, and turbulence playing secondary roles. The relative importance of these inputs differs across the mean, A, C, and E: spectral combinations are most prominent for C, while A and E share a driver set similar to the mean but with smaller overall sensitivities and increased response at extreme illumination and canopy states. These patterns are consistent across both test splits, and, given strong correlations among predictors, we interpret the SHAP results qualitatively as supporting evidence rather than as exact attributions.

Fifth, systematic hyperparameter sweeps reveal broad plateaus of good behavior rather than isolated peaks. Within the explored ranges of the number of mixture components, prior scale, and regularization strengths, performance and ACE patterns vary smoothly, and the final configuration is chosen from a region of near-optimal validation performance. The main ACE conclusions remain stable throughout this high-performing region, suggesting that the decomposition is reasonably robust to hyperparameter choices rather than being delicately tuned to a single setting.

Finally, structural ablations on feature families and Bayesian depth provide strong sanity checks. Removing spectral or meteorological features degrades predictive metrics and yields sharp but physically interpretable shifts in the ACE decomposition, especially in the compositional component. Increasing Bayesian depth beyond one layer leads to training instability and pathological decompositions. These

findings support the choice of a Bayesian last hidden layer and strongly support the interpretation that the compositional component is capturing meaningful ambiguity between distinct physical regimes rather than merely absorbing numerical artifacts.

Overall, the results in this chapter show that, for the multi-site AmeriFlux NEE dataset considered here, the BMDN and the ACE decomposition can be ported from an idealized synthetic setting to complex carbon flux data without losing interpretability. Even though we do not have ground truth for A, C, and E in the real data case, the decomposition behaves in ways that are consistent with both the theory and our physical understanding of ecosystem carbon fluxes. This is the central empirical claim of the thesis.

CHAPTER VII
CONCLUSION AND FUTURE WORK

7.1 Summary of contributions

This thesis proposed a principled framework for decomposing predictive uncertainty in mixture density networks into three components: *aleatoric* (A), *compositional* (C), and *epistemic* (E) uncertainty. Building on a Bayesian last-layer MDN, I derived the ACE decomposition from first principles, showed how it follows from the law of total variance, and clarified its relationship to more traditional two-way splits into aleatoric and epistemic uncertainty Kendall and Gal (2017); Walker et al. (2003). In particular, the work distinguishes genuinely irreducible noise (A) from multivalued structure in the conditional distribution (C), and from parameter uncertainty due to limited data (E).

In synthetic experiments with controlled ground truth, the ACE terms behaved consistently with their intended meaning. When the data were dominated by noise, the decomposition correctly assigned most of the variance to A. When the mapping from inputs to targets was deliberately made multimodal, C increased while A remained stable. When the training set was thinned or restricted in space, E increased in the under-sampled regions, while A and C remained essentially unchanged. These toy examples provided a first sanity check that the decomposition is identifiable and responds selectively to the mechanisms it is designed to represent.

Using a Bayesian MDN trained on AmeriFlux–Landsat NEE data, I then showed that the same structure carries over to a realistic geophysical setting that is relevant for natural climate solutions and land-sector mitigation Chausson et al. (2020); Griscom et al. (2017); IPCC (2023); Seddon et al. (2020); Smith et al. (2024). At the level of land-cover classes, croplands, mixed forests, and wetlands exhibited relatively high RMSE and were dominated by large aleatoric and compositional standard deviations, while grasslands and open shrublands had smaller A and much lower C and E despite similar R^2 in some splits. Beyond predictive accuracy, the ACE decomposition supports concrete decisions about where to improve measurements, add covariates, or deploy new towers by distinguishing noise-dominated, multi-regime, and data-scarce regimes.

Across hyperparameter sweeps, feature ablations, and Bayesian depth ablations, the ACE decomposition provided a coherent picture. Increasing the number of mixture components improved proper scoring rules and shifted variance from A to

C, as expected when multimodality is better captured. Removing key covariates inflated A and especially C, while leaving E relatively modest, pointing to missing features rather than data scarcity as the main driver of compositional uncertainty. SHAP analyses for the predictive mean and each ACE component link predictive and uncertainty behavior to physically interpretable drivers, primarily radiation and canopy state, and thereby help to identify potential interventions at the site or network scale.

Taken together, these results support the central claim of the thesis: decomposing predictive uncertainty into aleatoric, compositional, and epistemic components is both technically feasible and practically useful for carbon flux prediction. It provides insight that is not visible from a single predictive variance or from standard metrics such as RMSE and R^2, and it suggests different actions depending on which component dominates.

7.2 Answers to the research questions

The three research questions posed in Chapter I can be answered as follows.

RQ1: Accuracy and scalability of the MDN. Across synthetic experiments and AmeriFlux applications, the Bayesian Mixture Density Network produced accurate flux estimates with performance comparable to or better than the simpler probabilistic baselines considered here, while maintaining good behavior under temporal and spatial shift. On the synthetic problem, the BMDN matched the RMSE and R^2 of a heteroscedastic Gaussian regressor and a deterministic MDN while substantially improving NLL and CRPS. On AmeriFlux–Landsat data, the final model achieved $R^2 \approx 0.67$ on validation, $R^2 \approx 0.70$ on `test_future`, and $R^2 \approx 0.63$ on `test_site`, with RMSE in the range 2.8–3.1 across test splits. Conditioning on tower observations and applying the model to unsampled pixels provides a practical path to spatially scalable flux estimates.

RQ2: Effect of three-way uncertainty modeling and comparison to baselines. The ACE decomposition improved interpretability by separating noise-dominated regimes from multiregime structure and data-scarce regions. In the synthetic problem, controlled manipulations of noise, multimodality and data sparsity produced selective changes in A, C and E consistent with their intended meanings, and the BMDN improved test NLL from about -0.69 for a single-Gaussian baseline to around -1.52 at similar RMSE while maintaining well-behaved coverage. On

AmeriFlux data, calibrated predictive distributions and stable ACE patterns across splits and land-cover classes show that the three-component modeling does not compromise, and in some cases improves, calibration relative to simpler uncertainty baselines. At the same time, the experiments highlight that ACE is a model-based decomposition: different architectural or prior choices can shift variance between A, C, and E even when aggregate scores remain similar.

RQ3: Contribution of ancillary environmental drivers. Feature ablations on the real-data experiments showed that including meteorological and other ancillary drivers materially improves both point prediction and uncertainty decomposition. Removing these drivers degraded RMSE, R^2 and proper scoring rules and increased especially the compositional component, indicating that ancillary variables are essential for disambiguating regimes that cannot be separated using spectral information alone. Removing meteorological drivers was particularly damaging, with R^2 collapsing toward 0.25 and much larger A and C. In contrast, the full feature set achieved the best NLL and CRPS on both `test_future` and `test_site` and produced ACE patterns that align with physical expectations. Together with SHAP analyses, these ablations suggest that the chosen driver set is informative but not exhaustive, and that missing covariates tend to manifest as elevated compositional uncertainty.

RQ4: Decision support. The ACE decomposition provides concrete guidance for measurement design, feature collection, and deployment in natural climate solutions and carbon markets. High A highlights regimes where better instrumentation, post-processing, or footprint handling are needed; high C points to missing or insufficient covariates and motivates adding new sensors or remote-sensing products; high E identifies data-scarce regimes where additional towers or campaigns would most reduce uncertainty. The synthesis in Appendix C demonstrates how these patterns can be turned into a practical decision framework for prioritising towers, sensors, and credit discounting strategies.

7.3 Limitations

Despite these encouraging results, several limitations remain:

- **Approximate Bayesian inference.** In this work, the Bayesian treatment is restricted to the last layer and implemented via stochastic variational inference with an `AutoLowRankMultivariateNormal` guide, i.e. a low-rank-plus-diagonal Gaussian approximation to the weight posterior. This choice keeps training

and prediction computationally feasible, but it cannot represent multimodal or strongly non-Gaussian posteriors, and may therefore underestimate or distort epistemic uncertainty in highly non-linear or weakly constrained regions of parameter space.

- **Single-task, single-scale focus.** The work concentrates on half-hourly NEE at the flux-tower scale, using a single architecture and a fixed set of satellite and meteorological drivers. Extensions to multi-output settings (e.g., NEE, LE, H), other spatial scales, or alternative backbone architectures are not explored.
- **Limited spatiotemporal structure.** The model is essentially i.i.d. in time and space: spatial correlation, temporal autoregression, and cross-site dependence enter only via shared parameters, not through an explicit stochastic process model.
- **Footprint representation.** Flux-tower footprints and sub-pixel heterogeneity are only handled implicitly via covariates. The model does not yet incorporate explicit footprint weights or dynamic source areas in the training objective.
- **Baseline coverage.** The empirical evaluation compares BMDN primarily against internal baselines (a single-Gaussian heteroscedastic regressor and a deterministic MDN) rather than against the full ecosystem of probabilistic deep-learning and flux upscaling methods (for example, deep ensembles or FLUXCOM-style products). This limits our ability to claim practical superiority; the present results are best viewed as demonstrating that ACE can be implemented without sacrificing standard metrics, not as a definitive benchmark against all alternatives.
- **Model dependence of ACE.** The ACE decomposition is inherently model-based: different architectures, priors, and inference schemes that fit the data comparably well can allocate variance differently between A, C, and E. Throughout this thesis we have therefore treated ACE as a diagnostic lens for comparing models, regions, and design choices, rather than as an absolute physical split of uncertainty.
- **Evaluation scope.** The experiments are limited to one data set and a regionally biased network of towers. How ACE behaves under more extreme domain shifts (e.g., truly novel biomes or future climates) remains an open question.

7.4 Future directions

Several directions follow naturally from this work:

- **Richer posterior approximations and architectures.** A first avenue is to improve how epistemic uncertainty is represented. This includes exploring deeper Bayesian treatments (beyond a single layer) with more robust inference schemes, structured variational families, or ensembles explicitly calibrated to match Bayesian posteriors Blundell et al. (2015); Gal and Ghahramani (2016); Lakshminarayanan et al. (2017). Combining ACE with architectures that better respect spatial and temporal structure, such as spatiotemporal transformers or graph-based models over tower networks, could further sharpen the distinction between data scarcity (E) and model misspecification (manifesting in A and C).
- **Spatiotemporal and multi-task extensions.** Extending the BMDN+ACE framework to jointly model multiple fluxes (NEE, latent and sensible heat, methane) and to include explicit temporal dependence would make the decomposition more informative for process-based interpretation. In a multi-task setting, one could ask whether the same inputs that drive high compositional uncertainty in NEE also drive ambiguous behavior in other fluxes, or whether epistemic uncertainty is shared across variables and sites.
- **Footprint-aware training.** A complementary line of future work is to integrate flux-tower footprint information directly into the training objective, building on recent FLUXNET and footprint advances Chu et al. (2021); Delwiche et al. (2024); Pastorello et al. (2020). Instead of treating each tower observation as arising from a single representative pixel, one could weight surrounding pixels by a footprint kernel, or condition the mixture components on footprint summaries. This would better align the model with the physical source area of the measurements, potentially reducing aleatoric variance that currently reflects unresolved sub-pixel heterogeneity. Combining footprint-aware training with ACE would allow a more explicit separation between uncertainty due to unresolved source structure and uncertainty due to missing covariates or data scarcity.
- **Decision support for NCS and carbon markets.** The decomposition suggests concrete interventions that align well with emerging needs in natural climate solutions and carbon markets Chausson et al. (2020); Griscom et al.

(2017); IPCC (2023); Seddon et al. (2020); Smith et al. (2024). High A points to noisy measurements and motivates investments in better instrumentation, calibration, and post-processing. High C indicates missing or insufficient covariates and suggests adding sensors for key drivers (e.g., soil moisture, water level, management state). High E highlights data-scarce regimes and motivates additional tower deployments or targeted campaigns. Future work could operationalize this logic into tools that prioritize new towers, new sensors, or new data products under budget constraints, and that translate component-wise uncertainty into the total uncertainty figures used for crediting and discounting.

- **Broader benchmarking and robustness studies.** Applying the BMDN+ACE framework to other regions, ecosystems, and data modalities (e.g., global networks, airborne campaigns, or high-resolution imagery) would test its robustness and generality. Future work could include hybrid models that combine tower networks, satellite observations and large-scale constraints from atmospheric inversions or reanalyses Hersbach et al. (2020); Jung et al. (2019, 2020); Zeng et al. (2020). Systematic ablation and stress-testing under synthetic shifts and extreme events could help identify when the decomposition remains trustworthy and when more sophisticated models or priors are required.

In summary, this thesis demonstrates that a carefully structured uncertainty decomposition can turn a mixture density network from a black-box predictor into a more interpretable tool for diagnosing model behavior and guiding data collection. The next step is to embed that decomposition into richer models, more realistic spatiotemporal settings, and explicit decision frameworks for monitoring and managing land carbon.

APPENDIX A
PROOFS AND TECHNICAL EXTENSIONS

A.1 Mixture-of-uniforms structure and weights

Proof of Proposition 3.1. Under (3.1), the joint density of (Y, η) is constant on the rectangle $[0,1] \times [-\delta, \delta]$. Conditioning on $X = x$ imposes the constraint $\eta = x - g(y)$, so y is feasible if and only if $|x - g(y)| \le \delta$. Therefore

$$p(y \mid x) \propto \mathbf{1}\{|x - g(y)| \le \delta\}.$$

On each monotone branch B_j, the preimage of the band $[x-\delta, x+\delta]$ is a single interval $I_j(x) = [\ell_j(x), r_j(x)]$. Since the density is constant on the feasible set, conditioning and normalizing yield a mixture of uniform distributions on the non-empty intervals. The weight is the interval length divided by the total feasible measure, that is $w_j(x) = L_j(x) \big/ \sum_i L_i(x)$. For the small δ expansion, fix an x in the interior of the image of branch j and let $y_j^\star$ solve $g(y_j^\star) = x$ on that branch. By the mean value theorem,

$$g\big(y_j^\star \pm h\big) \approx g(y_j^\star) \pm h\, g'(y_j^\star) \quad \text{for small } h,$$

so the bounds $x \pm \delta$ map to $y_j^\star \pm \delta/|g'(y_j^\star)|$ up to higher order terms, which gives $L_j(x) \approx 2\delta/|g'(y_j^\star)|$. □

A.2 Exact variance decomposition

Proof of Proposition 3.2. Let $Z \in \{1, \dots, J(x)\}$ index the branch that contains Y given $X = x$. Under Proposition 3.1, $Y \mid X = x, Z = j \sim \mathrm{Unif}(\ell_j, r_j)$ with mean μ_j and variance $v_j = L_j^2/12$. The law of total variance implies

$$\mathrm{Var}[Y \mid X = x] = \mathbb{E}\big[\mathrm{Var}(Y \mid X = x, Z) \mid X = x\big] + \mathrm{Var}\big(\mathbb{E}[Y \mid X = x, Z] \mid X = x\big).$$

The first term equals $\sum_j w_j v_j = A(x)$. The second equals $\sum_j w_j (\mu_j - \mu)^2 = C(x)$. This completes the proof. □

A.3 Practical epistemic uncertainty exceeds the oracle

Proof of Proposition 3.3. Let $m_{\theta,Z}(x) = \mathbb{E}[Y \mid x, \theta, Z]$. Consider the joint distribution over (θ, Z) given x and data $\mathcal{D}$. The variance decomposition yields

$$\mathrm{Var}_{\theta,Z}\big(m_{\theta,Z}(x)\big) = \mathbb{E}_Z\big[\,\mathrm{Var}_\theta\big(m_{\theta,Z}(x)\big)\,\big] + \mathrm{Var}_Z\big(\,\mathbb{E}_\theta[m_{\theta,Z}(x)]\,\big).$$

The second term is non negative, hence $\mathrm{Var}_{\theta,Z}(m_{\theta,Z}(x)) \ge \mathbb{E}_Z\big[\mathrm{Var}_\theta(m_{\theta,Z}(x))\big]$. If the oracle is branch aware, it conditions on Z and averages the per branch variance of the predictive mean. Identifying $\mathrm{Var}_\theta(m_{\theta,Z}(x))$ with $v_{\mathrm{epi},Z}(x)$ from (3.4) gives the claim. □

A.4 Generalization beyond uniform prior or noise

Assume Y has density p_Y on $[0,1]$ and η has density r on $\mathbb{R}$, independent of Y. Then for any measurable set $A \subset [0,1]$,

$$\mathbb{P}(Y \in A \mid X = x) \;\propto\; \int_A p_Y(y)\, r\big(x - g(y)\big)\, dy.$$

Let Z be a branch label with support on the branches that intersect the feasible set. Define unnormalized weights

$$\tilde{w}_j(x) \;=\; \int_{\ell_j(x)}^{r_j(x)} p_Y(y)\, r\big(x - g(y)\big)\, dy, \qquad w_j(x) = \frac{\tilde{w}_j(x)}{\sum_i \tilde{w}_i(x)}.$$

Within branch j, the conditional density of Y is

$$p_j(y \mid x) = \frac{p_Y(y)\, r\big(x - g(y)\big)\, \mathbf{1}\{y \in I_j(x)\}}{\tilde{w}_j(x)}.$$

The within component mean and variance become

$$\mu_j(x) = \int_{I_j(x)} y\, p_j(y \mid x)\, dy, \qquad v_j(x) = \int_{I_j(x)} \big(y - \mu_j(x)\big)^2\, p_j(y \mid x)\, dy.$$

The decomposition $\mathrm{Var}[Y \mid X = x] = \sum_j w_j v_j + \sum_j w_j(\mu_j - \mu)^2$ still holds, but closed forms (3.2) and (3.3) do not. The uniform case recovers the simple length based weights and uniform moments.

A.5 From theory to estimation

In practice we work with a parametric or Bayesian nonparametric model for $p(y \mid x, \theta)$. The uncertainty mapping is:

- $A(x)$ maps to the within component dispersion of $p(y \mid x, \theta)$ at fixed θ, then averaged over the posterior.
- $C(x)$ maps to the between component dispersion of component means around the mixture mean at fixed θ, then averaged over the posterior.
- $E(x)$ is the posterior variance of the predictive mean $\mathbb{E}[Y \mid x, \theta]$.

Calibration diagnostics compare estimated A, C, E against the ground truth in Chapter III. The oracle in (3.4) sets a lower bound for epistemic dispersion when the component identity is revealed.

APPENDIX B
EXTENDED CARBON FLUX RESULTS

Table B.7. Per IGBP performance and ACE decomposition on the training split. A_sd, C_sd, and E_sd denote the mean aleatoric, compositional, and epistemic standard deviations of the predictive distribution within each IGBP class.

IGBP	RMSE	R^2	count	A_sd	C_sd	E_sd
BSV	0.99	0.0233	1,213	0.94	0.38	0.31
CRO	3.63	0.7051	327,201	2.62	1.61	0.93
CSH	1.36	0.8566	231,984	1.61	0.77	0.54
CVM	2.73	0.6693	16,059	2.68	1.28	0.96
DBF	2.85	0.7612	378,327	2.25	1.13	0.75
DNF	2.02	0.0547	64	2.41	1.32	0.77
EBF	3.61	0.6365	22,586	3.31	1.60	1.26
ENF	2.52	0.7171	712,264	2.14	1.04	0.72
GRA	2.38	0.6689	417,569	1.74	0.89	0.59
MF	3.49	0.7811	50,719	2.95	1.44	1.00
OSH	1.40	0.6144	115,386	1.52	0.61	0.51
SAV	1.29	0.6762	113,709	1.44	0.64	0.47
SNO	2.23	0.2152	25,423	1.44	0.53	0.45
WAT	2.75	0.0384	66,268	1.72	0.76	0.54
WET	2.46	0.7793	428,228	2.18	1.04	0.72
WSA	1.72	0.6711	228,881	1.47	0.68	0.49

Table B.8. Per IGBP performance and ACE decomposition on the validation split. A_sd, C_sd, and E_sd denote the mean aleatoric, compositional, and epistemic standard deviations of the predictive distribution within each IGBP class.

IGBP	RMSE	R^2	count	A_sd	C_sd	E_sd
BSV	0.63	0.0055	624	0.96	0.43	0.32
CRO	4.57	0.5368	66,747	2.68	1.67	0.97
CSH	1.57	0.8300	49,580	1.65	0.80	0.55
CVM	2.53	0.7286	2,133	2.88	1.52	1.02
DBF	3.22	0.6862	78,825	2.37	1.20	0.81
DNF	2.37	0.2783	376	2.62	1.37	0.82
EBF	3.87	0.5796	4,297	3.26	1.61	1.20
ENF	2.64	0.7104	155,666	2.21	1.07	0.75
GRA	2.53	0.5994	95,846	1.74	0.90	0.59
MF	3.40	0.7766	9,762	2.86	1.40	0.95
OSH	1.38	0.5241	24,461	1.52	0.61	0.49
SAV	1.46	0.5953	22,078	1.50	0.69	0.48
SNO	2.76	0.1050	3,342	1.55	0.56	0.44
WAT	2.65	-0.0138	14,000	1.63	0.70	0.49
WET	2.40	0.7650	101,806	2.04	0.96	0.68
WSA	1.82	0.6438	53,575	1.52	0.72	0.50

APPENDIX C
SYNTHESIS AND DECISION FRAMEWORK

This chapter pulls together the main findings of the thesis and turns them into a simple decision framework. First, I summarize how the toy and real-data experiments jointly support the interpretation of aleatoric, compositional, and epistemic (ACE) uncertainty in the Bayesian Mixture Density Network (BMDN). I then describe how ACE can guide practical actions in measurement design, feature collection, and deployment, especially in the context of natural climate solutions (NCS) and carbon markets Chausson et al. (2020); Griscom et al. (2017); IPCC (2023); Seddon et al. (2020); Smith et al. (2024). Finally, I outline the main threats to validity and how they qualify the conclusions.

C.1 Cross-chapter synthesis

C.1.1 From total variance to ACE. The thesis moves from a single "total predictive variance" to a three-way split tailored to mixture models. As defined in Chapter III, we decompose predictive variance into three components; within the BMDN these correspond to:

- **Aleatoric (A)**: mixture-weighted within-component variance (noise and sub-grid variability within a regime),
- **Compositional (C)**: between-component variance of component means (multi-regime structure at fixed covariates),
- **Epistemic (E)**: variance of the predictive mean across posterior weight samples (uncertainty due to limited or unrepresentative data), in line with broader distinctions between aleatoric and epistemic uncertainty in machine learning Aven (2010); Kendall and Gal (2017); Walker et al. (2003).

Earlier deterministic MDNs effectively bundled A and C, labeling the between-component term as "epistemic". Within ACE, that term is more naturally interpreted as compositional: it reflects stable mixture structure that cannot be removed without changing the model, whereas epistemic uncertainty should, at least in principle, shrink with more data or better priors.

C.1.2 Toy experiments: controlled checks. The synthetic experiments showed that ACE behaves as intended when one mechanism is varied at a time:

- *Noise-only changes*: increasing observation noise raises A, with C and E remaining small and stable.

– *Multi-modal mappings*: introducing more modes or separating them more clearly moves variance from A into C, while E stays small, confirming that C is capturing genuine multivaluedness rather than lack of data.
– *Data sparsity and shift*: removing training data or shifting covariates locally increases E in those regions, with A and C largely unchanged where data remain dense.

Hyperparameter sweeps over the number of mixture components, the prior scale, and regularization showed that ACE is numerically stable over a broad "sensible" region: moving from $K = 1$ to $K \geq 2$ improves proper scoring rules and gradually reallocates variance from A to C, while E remains almost flat; excessively strong priors inflate E and hurt all metrics, while moderate-to-weak priors yield lower E and stable A and C.

C.1.3 Real-data experiments: patterns across ecosystems and sites. On AmeriFlux data, the same qualitative patterns reappear in a much messier setting:

– *By land cover*, ecosystems with similar R^2 can have very different uncertainty profiles. Croplands, wetlands and some forests show large A and C; grasslands and open shrublands often have similar or better R^2 but much smaller C and E. Thus similar "accuracy" can hide very different underlying ambiguity.
– *By tower*, sites like MX-PMm (mangrove) and US-Bi1 (intensively managed cropland) combine high RMSE with large A and C and elevated E, consistent with hydrological and management-driven regime shifts. Semi-arid shrublands such as US-Whs and US-Ses have low RMSE, very small C and E, and near-zero fluxes, matching the expectation of simple, low-amplitude behavior. Intermediate sites such as permafrost peatlands sit between these extremes.

A desirable property is also confirmed: where the model has high RMSE, it usually predicts high $A+C+E$, and sites with many observations tend to have lower E. That is, the model is not confidently wrong in the hardest regimes.

C.1.4 Drivers of mean and ACE. SHAP analyses link ACE back to physical drivers Covert et al. (2021); Lundberg and Lee (2017); Shapley (1953):

– *Mean NEE* is dominated by radiation and spectral bands, with meteorology and turbulence as secondary but interpretable contributors.
– *Aleatoric (A)* is strongly influenced by turbulence and high radiation, matching expectations about measurement uncertainty, footprint variability, and high-frequency processes.

– *Compositional (C)* is driven mainly by spectral combinations, supporting the idea that unresolved flux regimes arise when similar reflectance corresponds to different underlying states (e.g. management or hydrology).
– *Epistemic (E)* is largest in rare or extreme combinations of radiation and spectra, highlighting under-sampled regions of feature space.

C.1.5 Calibration and consistency. Reliability curves and PIT histograms indicate that the BMDN's predictive distributions are well calibrated on both temporal and spatial test splits. Empirical quantiles closely track nominal quantiles; PIT histograms are close to uniform, with only mild deviations. Together with good proper scores, this justifies treating A, C and E as meaningful diagnostics rather than artifacts of a badly misspecified likelihood.

Across chapters, the story is consistent: A responds to noise and fine-scale heterogeneity, C responds to unresolved multi-regime structure, and E responds to data scarcity and shift. The rest of the chapter uses this to build a decision framework.

C.2 Decision guide

C.2.1 Preconditions. Before using ACE for decisions:

1. **Check calibration.** If reliability and PIT diagnostics are poor, fix the model (architecture, likelihood, priors, training) before interpreting A, C or E.
2. **Stay in a stable hyperparameter regime.** Avoid settings where the sweeps showed pathological behavior (e.g. overly strong priors, too many Bayesian layers, or $K = 1$ with poor scores).

C.2.2 Step 1: Identify what dominates. For a land-cover class, tower, or spatial location, define the total predictive variance

$$V_{\text{tot}} = A + C + E$$

and the corresponding fractions

$$f_A = \frac{A}{V_{\text{tot}}}, \quad f_C = \frac{C}{V_{\text{tot}}}, \quad f_E = \frac{E}{V_{\text{tot}}}.$$

Then classify:

– A**-dominated**: f_A large, f_C and f_E small → noisy but essentially unimodal.
– C**-dominated**: C comparable to A and much larger than E → multiple regimes under similar covariates.
– E**-elevated**: E unusually large relative to the dataset average → data-scarce or out-of-distribution region.

C.2.3 Step 2: Match actions to A, C, and E. The main conclusions can be summarized as follows:

1. **If A is high, focus on measurements and representation.**
   Large aleatoric uncertainty suggests that, even given the covariates, residual variability is high. The most effective actions are:
   - improve sensors and maintenance,
   - use better post-processing (energy-balance closure, quality control),
   - improve footprint modeling or temporal aggregation to reduce representativeness error.
2. **If C is high, add regime-resolving covariates.**
   Large compositional uncertainty indicates unresolved regimes: the same observed state corresponds to several plausible flux outcomes. The main actions are:
   - add new sensors or covariates (e.g. soil moisture, water table, management records),
   - incorporate additional remote-sensing products (e.g. thermal, SAR, SIF),
3. **If E is high, collect more or better-targeted data.**
   Elevated epistemic uncertainty means predictions are sensitive to the choice of weights and priors. Key levers are:
   - deploy new towers or campaigns in high-E regions,
   - refine priors and Bayesian depth to avoid both under- and over-confidence, using epistemic maps to prioritize data acquisition Gal and Ghahramani (2016); Lakshminarayanan et al. (2017).

In carbon accounting and NCS, total uncertainty still drives discounts and buffers IPCC (2023); Smith et al. (2024). What ACE adds is guidance on *how* to reduce that total: better instruments and processing when A dominates, new features and regime labels when C dominates, and targeted data collection when E dominates.

C.2.4 Suggested flowchart. A decision flowchart can make this procedure explicit. Figure C.18 illustrates one possible design:

1. Is the model calibrated? If not, improve modeling.
2. Where is V_{tot} (and RMSE) too large relative to project needs?
3. For those locations/classes/towers, is A, C, or E dominant?
4. Apply the corresponding interventions (measurement, features, data).
5. Re-train, re-evaluate calibration and ACE, and iterate as needed.

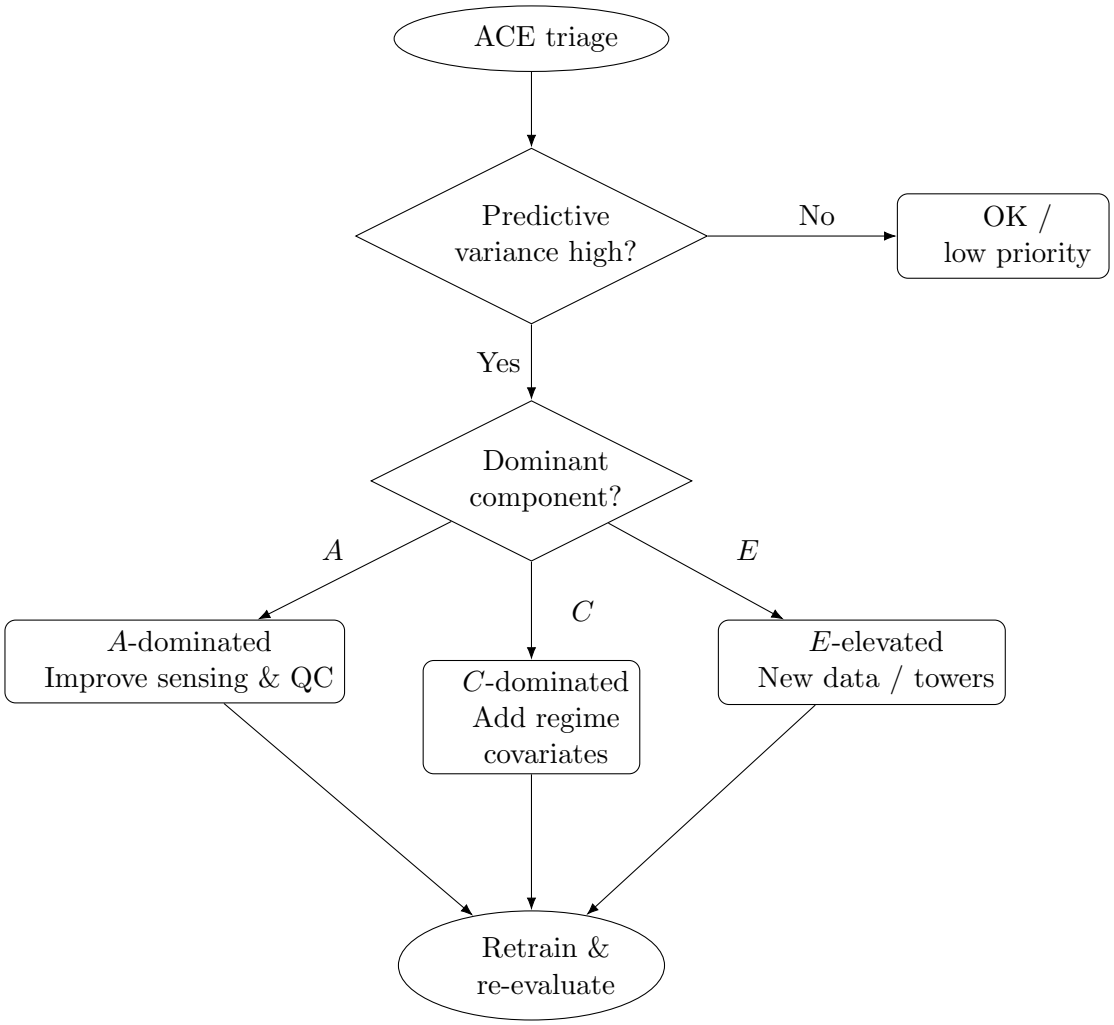


Figure C.18. ACE-based triage for deciding how to reduce predictive uncertainty.

C.3 Threats to validity

Finally, several limitations qualify these conclusions.

C.3.1 Modelling.

- **Approximate inference**: Epistemic uncertainty depends on the variational approximation and prior choices; different approximations could yield different E with similar predictive metrics.
- **Hyperparameter sensitivity**: ACE depends on K, network depth, and which layers are Bayesian. Depth ablations showed regimes where E collapses and A/C inflate, even with worse R^2.

C.3.2 Data and covariates.

- **Flux uncertainty and processing**: Eddy-covariance data include measurement error, energy-balance non-closure, gap-filling artifacts, and unmodeled sub-mesoscale processes; some of what appears as A is really "uncertainty about the labels".
- **Footprint mismatch and sub-grid heterogeneity**: The tower footprint and satellite pixel rarely match perfectly. Heterogeneity is currently absorbed into A and C rather than modelled explicitly.
- **Missing drivers**: Soil moisture, water table, management, species composition, and disturbance history are only partly represented. High C often reflects these missing covariates, so the decomposition is always relative to the chosen feature set.
- **Coverage and transferability**: Towers are unevenly distributed across climates and land covers. E in under-sampled regions is therefore conditional on the existing network.

C.3.3 Evaluation and usage.

- **Split design**: Temporal and spatial splits are realistic but not exhaustive; behavior under future climate or land-cover change is not explicitly tested.
- **Aggregate metrics**: Global RMSE, R^2, NLL, and CRPS hide heterogeneity; rare events and seasonal extremes are not separately analyzed.
- **Model dependence of ACE**: The ACE decomposition is inherently model-based. Different architectures, priors, or inference schemes that fit the data similarly well can yield different allocations of variance between A, C, and E. For this reason ACE should be interpreted comparatively, across models, regions, or design choices, rather than as an absolute decomposition of physical uncertainty.
- **Regulatory context**: Current carbon standards operate on total uncertainty, not component-wise splits. ACE is best viewed as a diagnostic layer to inform internal decisions and measurement design, not as a direct replacement for existing rules.

Despite these caveats, the results across synthetic and real experiments are mutually consistent: ACE is numerically stable in sensible configurations, aligns with controlled manipulations, and produces physically interpretable patterns across sites and land covers. This supports its use as a practical tool for understanding

and managing uncertainty in carbon flux prediction, and for planning targeted improvements in measurement, modeling, and data collection.

REFERENCES CITED

Aven, T. (2010). *Risk analysis: Assessing uncertainties beyond expected values and probabilities*. Chichester, UK: John Wiley & Sons.

Baldocchi, D. (2003). Assessing the eddy covariance technique for evaluating carbon dioxide exchange rates of ecosystems: Past, present and future. *Global Change Biology*, *9*(4), 479–492.

Beer, C., et al. (2010). Terrestrial gross carbon dioxide uptake: global distribution and covariation with climate. *Science*, *329*(5993), 834–838. doi: 10.1126/science.1184984

Bishop, C. M. (1994). *Mixture density networks* (Tech. Rep.). Birmingham, UK: Neural Computing Research Group, Aston University.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York, NY: Springer.

Blundell, C., Cornebise, J., Kavukcuoglu, K., & Wierstra, D. (2015). Weight uncertainty in neural networks. In *Proceedings of the 32nd international conference on machine learning* (pp. 1613–1622).

Bosilovich, M. G., Robertson, F. R., & Takacs, L. (2015). *Merra-2: Initial evaluation of the climate* (Tech. Rep. Nos. NASA/TM-2015-104606, Vol. 43). NASA Goddard Space Flight Center.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. doi: 10.1023/A:1010933404324

Bröcker, J., & Smith, L. A. (2007). Increasing the reliability of reliability diagrams. *Weather and Forecasting*, *22*(3), 651–661. doi: 10.1175/WAF993.1

Chausson, A., Turner, B., Seddon, N., Chabaneix, N., Girardin, C. A. J., Kapos, V., ... Woroniecki, S. (2020). Mapping the effectiveness of nature-based solutions for climate change adaptation. *Global Change Biology*, *26*(11), 6134–6155.

Chu, H., Luo, X., Ouyang, Z., Chan, W. S., Dengel, S., Podgrajsek, E., ... others (2021). Representativeness of eddy-covariance flux footprints for areas surrounding ameriflux sites. *Agricultural and Forest Meteorology*, *301–302*, 108350. doi: 10.1016/j.agrformet.2021.108350

Covert, I., Lundberg, S. M., & Lee, S.-I. (2021). Explaining by removing: A unified framework for model explanation. *Journal of Machine Learning Research*, *22*(209), 1–90.

Delwiche, K. B., Papale, D., Chu, H., Arriga, N., Baldocchi, D., et al. (2024). Charting the future of the fluxnet network. *Bulletin of the American Meteorological Society*, *105*(3), E531–E555.

Depeweg, S., Hernández-Lobato, J. M., Doshi-Velez, F., & Udluft, S. (2018). Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. *Proceedings of the 35th International Conference on Machine Learning Workshop on Reliable Machine Learning in the Wild*.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *13*(3), 253–263. doi: 10.1080/07350015.1995.10524599

Dulal, A., & Searcy, J. (2025). Uncertainty-aware carbon flux estimation from multispectral landsat imagery using mixture density networks. In *Iclr 2025 workshop on tackling climate change with machine learning*. Retrieved from `https://www.climatechange.ai/papers/iclr2025/15`

Friedlingstein, P., Jones, M. W., O'Sullivan, M., Andrew, R. M., Bakker, D. C. E., Hauck, J., ... others (2023). Global carbon budget 2023. *Earth System Science Data*, *15*(12), 5301–5369.

Gal, Y., & Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd international conference on machine learning* (pp. 1050–1059).

Gneiting, T., Balabdaoui, F., & Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B*, *69*(2), 243–268.

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359–378.

Griscom, B. W., Adams, J., Ellis, P. W., Houghton, R. A., Lomax, G., Miteva, D. A., ... others (2017). Natural climate solutions. *Proceedings of the National Academy of Sciences*, *114*(44), 11645–11650.

Haya, B. K., Cullenward, D., Strong, A. L., Grubert, E., Heilmayr, R., Sivas, D. A., & Wara, M. (2020). Managing uncertainty in carbon offsets: Insights from california's standardized approach. *Climate Policy*, *20*(9), 1112–1126. doi: 10.1080/14693062.2020.1781035

Hepp, T., Zierk, J., & Seitz, S. (2022). Mixture density networks for the indirect estimation of reference intervals. *BMC Bioinformatics*, *23*(1), 57.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., ... others (2020). The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999–2049.

Herzallah, R., & Lowe, D. (2004). A mixture density network approach to modelling and exploiting uncertainty in nonlinear control problems. *Engineering Applications of Artificial Intelligence*, *17*(2), 145–158. doi: 10.1016/j.engappai.2004.02.001

Hjorth, L. U., & Nabney, I. T. (1999). Regularisation of mixture density networks. In *Proceedings of the 9th international conference on artificial neural networks* (pp. 521–526). doi: 10.1049/cp:19991162

IPCC. (2023). *Climate change 2023: Synthesis report. contribution of working groups i, ii and iii to the sixth assessment report of the intergovernmental panel on climate change*. Geneva, Switzerland: Intergovernmental Panel on Climate Change. doi: 10.59327/IPCC/AR6-9789291691647

Jung, M., Koirala, S., Weber, U., Ichii, K., Gans, F., Camps-Valls, G., ... Tramontana, G. (2019). The fluxcom ensemble of global land–atmosphere energy fluxes. *Scientific Data*, *6*, 74.

Jung, M., Schwalm, C. R., Migliavacca, M., Walther, S., Camps-Valls, G., Koirala, S., ... others (2020). Scaling carbon fluxes from eddy covariance sites to the globe: Synthesis and evaluation of the fluxcom approach. *Biogeosciences*, *17*(5), 1343–1365.

Justice, C. O., Townshend, J. R. G., Vermote, E. F., Masuoka, E., Wolfe, R. E., Saleous, N., ... Morisette, J. T. (2002). An overview of modis land data processing and product status. *Remote Sensing of Environment*, *83*(1–2), 3–15.

Justice, C. O., Vermote, E., Townshend, J. R. G., Defries, R., Roy, D. P., Hall, D. K., ... others (1998). The moderate resolution imaging spectroradiometer (modis): Land remote sensing for global change research. *IEEE Transactions on Geoscience and Remote Sensing*, *36*(4), 1228–1249.

Kendall, A., & Gal, Y. (2017). What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems* (Vol. 30).

Kingma, D. P., & Welling, M. (2014). Auto-encoding variational bayes. In *Proceedings of the 2nd international conference on learning representations.*

Kljun, N., Calanca, P., Rotach, M. W., & Schmid, H. P. (2015). A simple two-dimensional parameterisation for flux footprint prediction (ffp). *Geoscientific Model Development*, *8*(11), 3695–3713. doi: 10.5194/gmd-8-3695-2015

Kormann, R., & Meixner, F. X. (2001). An analytical footprint model for non-neutral stratification. *Boundary-Layer Meteorology*, *99*(2), 207–224. doi: 10.1023/A:1018991015119

Kuleshov, V., Fenner, N., & Ermon, S. (2018). Accurate uncertainties for deep learning using calibrated regression. In *Proceedings of the 35th international conference on machine learning* (pp. 2801–2809).

Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in neural information processing systems* (Vol. 30).

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in neural information processing systems* (Vol. 30).

Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research*, *7*, 983–999.

Men, Z., Yee, E., Lien, F.-S., Wen, D., & Chen, Y. (2016). Short-term wind speed and power forecasting using an ensemble of mixture density neural networks. *Renewable Energy*, *87*, 203–211. doi: 10.1016/j.renene.2015.10.019

Morfitt, R., Barsi, J. A., Levy, R., Markham, B., Micijevic, E., Ong, L., ... Vanderwerff, K. (2015). Landsat-8 operational land imager (oli) radiometric performance on-orbit. *Remote Sensing*, *7*(2), 2208–2237.

NASA Landsat Science Team. (n.d.). *Landsat 8: Instruments and mission facts.* https://landsat.gsfc.nasa.gov/satellites/landsat-8/. (Accessed 2025-10-30)

Neal, R. M. (1996). *Bayesian learning for neural networks.* New York, NY: Springer.

Papale, D. (2020). Ideas and perspectives: Enhancing the impact of the fluxnet network of eddy covariance sites. *Biogeosciences*, *17*(22), 5587–5598. doi: 10.5194/bg-17-5587-2020

Pastorello, G., Trotta, C., Canfora, E., Chu, H., Christianson, D., Cheah, Y. W., ... others (2020). The fluxnet2015 dataset and the oneflux processing pipeline for eddy covariance data. *Scientific Data*, *7*, 225.

Roy, D. P., Wulder, M. A., Loveland, T. R., Woodcock, C. E., Allen, R. G., Anderson, M. C., ... others (2014). Landsat-8: Science and product vision for terrestrial global change research. *Remote Sensing of Environment*, *145*, 154–172.

Saranathan, A. M., et al. (2024). Assessment of advanced neural networks for the dual estimation of water quality indicators and their uncertainties. *Frontiers in Remote Sensing*, *5*, 1383147. doi: 10.3389/frsen.2024.1383147

Schaaf, C. B., Gao, F., Strahler, A. H., Lucht, W., Li, X., Tsang, L., ... others (2002). First operational brdf, albedo and nadir reflectance products from modis. *Remote Sensing of Environment*, *83*(1–2), 135–148.

Schmid, H. P. (2002). Footprint modeling for vegetation–atmosphere exchange studies: A review and perspective. *Agricultural and Forest Meteorology*, *113*(1–4), 159–183.

Searcy, J., Dulal, A., Bridgham, S., Cordes, A., Aoki, L., Bohannan, B., ... Silva, L. C. (2025). A footprint-aware, high-resolution approach for carbon flux prediction across diverse ecosystems. *arXiv preprint arXiv:2512.01917*.

Seddon, N., Chausson, A., Berry, P., Girardin, C. A. J., Smith, A., & Turner, B. (2020). Understanding the value and limits of nature-based solutions to climate change and other global challenges. *Philosophical Transactions of the Royal Society B*, *375*(1794), 20190120.

Shapley, L. S. (1953). A value for n-person games. In H. W. Kuhn & A. W. Tucker (Eds.), *Contributions to the theory of games, volume ii* (pp. 307–317). Princeton, NJ: Princeton University Press.

Smith, S., Tighe, M., Fuss, S., Canadell, J., Minx, J., et al. (2024). *The state of carbon dioxide removal 2024.* Report. Retrieved from `https://cdr-stateoftheart.org` (Accessed: 2025-11-29)

Teo, H. C., Tan, N. H. L., Zheng, Q., Lim, A. J. Y., Sreekar, R., Chen, X., ... Koh, L. P. (2023). Uncertainties in deforestation emission baseline methodologies and implications for carbon markets. *Nature Communications*, *14*(8277), 1–10. doi: 10.1038/s41467-023-44127-9

Tramontana, G., Jung, M., Schwalm, C. R., Ichii, K., Camps-Valls, G., Rätch, G., ... others (2016). Predicting carbon dioxide and energy fluxes across global fluxnet sites with regression algorithms. *Biogeosciences*, *13*(14), 4291–4313.

Unni, R., Yao, K., & Zheng, Y. B. (2020). Deep convolutional mixture density network for inverse design of layered photonic structures. *ACS Photonics*, *7*(10), 2703–2712.

(USGS), U. G. S., Aeronautics, N., & (NASA), S. A. (2024). *Landsat 8 (landsat data continuity mission)*. NASA Landsat Science Website. Retrieved from `https://landsat.gsfc.nasa.gov/satellites/landsat-8/` (`https://landsat.gsfc.nasa.gov/satellites/landsat-8/`)

Vermote, E., Justice, C., Claverie, M., & Franch, B. (2016). Preliminary analysis of the performance of the landsat 8 oli land surface reflectance product. *Remote Sensing of Environment*, *185*, 46–56.

Vesala, T., Kljun, N., Rannik, Ü., Rinne, J., Sogachev, A., Markkanen, T., ... Leclerc, M. Y. (2008). Flux and concentration footprint modelling: State of the art. *Environmental Pollution*, *152*(3), 653–666.

Walker, W. E., Harremoës, P., Rotmans, J., van der Sluijs, J. P., van Asselt, M. B. A., Janssen, P., & Krayer von Krauss, M. P. (2003). Defining uncertainty: A conceptual basis for uncertainty management in model-based decision support. *Integrated Assessment*, *4*(1), 5–17.

Wulder, M. A., Loveland, T. R., Roy, D. P., Crawford, C. J., Masek, J. G., Woodcock, C. E., ... others (2019). Current status of the landsat program, science, and applications. *Remote Sensing of Environment*, *225*, 127–147.

Zeng, N., Zhao, F., Collatz, G., Kalnay, E., Salawitch, R., West, T. O., ... others (2020). Global terrestrial carbon fluxes of 1999–2019 estimated by neural network constrained by co_2 observations. *Environmental Research Letters*, *15*(12), 124050.

Zhao, M., Heinsch, F. A., Nemani, R. R., & Running, S. W. (2005). Improvements of the modis terrestrial gross and net primary production global data set. *Remote Sensing of Environment*, *95*(2), 164–176.