Received: 28 November 2025 | Revised: 17 March 2026 | Accepted: 22 April 2026

DOI: 10.1002/qj.70219

Quarterly Journal of the Royal Meteorological Society RMetS

RESEARCH ARTICLE

Epistemic and aleatoric uncertainty quantification in weather and climate models

Laura A. Mansfield | **Hannah M. Christensen**

Atmospheric, Oceanic and Planetary Physics, University of Oxford, Oxford, UK

Correspondence
Laura A. Mansfield, Atmospheric, Oceanic and Planetary Physics, University of Oxford, Oxford, UK.
Email: laura.mansfield@physics.ox.ac.uk

Funding information
Schmidt Sciences; HORIZON EUROPE Climate, Energy and Mobility, Grant/Award Numbers: 101081383, 10049639; Leverhulme Trust

Abstract

Representing and quantifying uncertainty in physical parameterisations is a central challenge in weather and climate modelling, and approaches are often developed separately for different time-scales. Here, we introduce a unified framework for analysing uncertainty in parameterisations across weather and climate regimes. Using the Lorenz 1996 system as a testbed for simplified chaotic dynamics, we quantify uncertainties in a subgrid-scale parameterisation using a Bayesian neural network (BNN). This allows us to disentangle aleatoric uncertainty, arising from internal variability in the training data, and epistemic uncertainties, arising from poorly constrained parameters during training. At runtime, we sample uncertainties in line with stochastic approaches in weather models and perturbed-parameter methods in climate models. On weather time-scales, aleatoric uncertainty dominates, underscoring the value of stochastic parameterisations. On longer, climate time-scales and under changing forcings, accounting for both types of uncertainty is necessary for well-calibrated ensembles, with epistemic uncertainty widening the range of explored climate states, and aleatoric uncertainty promoting transitions between them. Constraining parameter uncertainty with short simulations reduces epistemic uncertainty and improves long-term model behaviour under perturbed forcings. This framework links concepts from machine learning with traditional uncertainty quantification in earth system modelling, offering a pathway towards seamless treatment of uncertainty in weather and climate prediction.

KEYWORDS

Bayesian neural networks, climate, earth system models, epistemic and aleatoric, machine learning, parameterisations, uncertainty quantification, weather

1 | INTRODUCTION

1.1 | General circulation models

General circulation models (GCMs) simulate the earth's atmosphere, ocean, land surface, and sea ice. They consist of a dynamical core which solves the governing equations of fluid motion on a 3D grid for the earth, and physical parameterisations or closures, that represent unresolved subgrid-scale processes, such as radiation, convection, clouds and aerosol microphysics, atmospheric gravity waves, ocean eddies, and land surface processes (Christensen & Zanna, 2022). Parameterisations are typically based on simplified equations or empirical relationships and make several assumptions which can introduce a large source of uncertainty into GCM output.

Recently, machine learning (ML) and artificial intelligence (AI) approaches have gained popularity for learning subgrid-scale parameterisations (e.g., Behrens *et al.*, 2022; Christensen & Zanna, 2022; Heuer *et al.*, 2024; Rasp *et al.*, 2018; Souza *et al.*, 2020; Ukkonen & Chantry, 2025; Yu *et al.*, 2024; Yuval & O'Gorman, 2023). However, while uncertainty quantification (UQ) has long been a crucial aspect of parameterisation development (e.g., Berner *et al.*, 2017), it is often overlooked in ML-based approaches (Christensen *et al.*, 2024). This paper explores UQ methods that could be leveraged as we transition towards AI-enhanced GCMs. We examine the roles of epistemic (model) and aleatoric (data) uncertainty and how they fit in with the traditional view of uncertainties in GCMs. We advocate that UQ should remain an essential part of parameterisation development, regardless of whether it is physics-based or data-driven.

1.2 | Uncertainties in general circulation models

Understanding and communicating uncertainties are an essential part of weather and climate prediction (Gigerenzer *et al.*, 2005). Meteorological agencies, governments, energy providers, the agriculture sector, insurance companies, among others need to make decisions based on GCM output. For instance, public weather forecasts rely on probabilistic predictions (Gneiting & Katzfuss, 2014). These are especially important for low-probability but high-impact events, such as heavy rainfall and flooding (Cloke & Pappenberger, 2009). These may require advanced warnings or even government action to reduce the impact on society. On longer time-scales, GCMs are used to generate projections of future climate change projections that are essential for policy-makers developing long-term strategies for climate adaptation (e.g., Calvin *et al.*, 2023; Stainforth *et al.*, 2005). Furthermore, scientific studies that predict climate model response to different forcings, physics, or geography, must also consider potential uncertainties before making conclusions (e.g., Forster *et al.*, 2013; Murphy *et al.*, 2004).

1.2.1 | Weather forecasting

On weather prediction time-scales, uncertainty in the initial conditions become amplified due to the chaotic processes in the atmosphere. This is initial condition uncertainty and can be captured by perturbing the input state and running ensembles (Slingo & Palmer, 2011). There are also uncertainties associated with the model formulation, particularly in the choices and assumptions made when developing parameterisations. This is known as model uncertainty (Slingo & Palmer, 2011). Usually, this term refers to uncertainties in the structure and parameters of the model or parameterisation. When developing parameterisations, there is also the issue that for a given resolved state, there can be many possible unresolved states. This is the subgrid variability and can be dealt with through stochastic parameterisations (Berner *et al.*, 2017).

Deterministic parameterisations assume that, given resolved state variables, the tendencies from unresolved processes can be estimated with a deterministic function derived from the mean grid-box behaviour. This assumption is valid when there is a clear-scale separation between resolved and unresolved processes (Berner *et al.*, 2017; Christensen & Zanna, 2022; Palmer, 2019). However, as model resolution increases and more processes become partially resolved, this scale separation breaks down. In these 'grey zone' regimes, the processes are not explicitly resolved but too few subgrid-scale processes exist within a single grid cell to define the mean grid-box behaviour. For instance, this occurs at around the order of 10s–100s of km for atmospheric organised moist convection (Christensen & Zanna, 2022). Instead of taking the average over many subgrid-scale processes, we must sample a single realisation of that subgrid-scale process, creating stochastic parameterisations (Berner *et al.*, 2017). An example of implementing this in practice is the stochastically perturbed parameterisation tendency (SPPT) scheme, which combines a deterministic estimate of the most likely subgrid-scale tendency with a stochastic perturbation to represent its variability (often correlated in time and space, Buizza *et al.*, 1999).

1.2.2 | Climate change projections

In contrast, on climate time-scales we are generally interested in *climate change projections*, where we predict the climate response to a forcing. This is a boundary condition problem rather than an initial condition problem. Initial-condition uncertainty is not considered a large source of uncertainty on these time-scales (Slingo & Palmer, 2011), as we are interested in climate statistics. Furthermore, subgrid variability may be less pronounced due to the larger scale separation. Instead, the main three types of uncertainty are internal variability (which causes natural variations on decadal time-scales), model uncertainty (in model structure or parameters, which define how different models respond differently to the same forcing) and scenario uncertainty (our lack of knowledge about the future forcing itself) (Hawkins & Sutton, 2009). Internal variability can be captured by repeating climate forcing experiments with different initialisations for the ocean state to capture variations in the response that occurs on decadal time-scales. In contrast, the latter two types are more relevant on longer, centennial time-scales. Here, we will not consider scenario uncertainty, which requires running simulations under several possible future pathways, and instead focus on model uncertainty.

Similarly to NWP, climate model uncertainty arises from the different modelling choices that can be made during model development. One approach to represent model uncertainty is to consider multimodel ensembles from many different modelling centres (e.g., multimodel coupled model intercomparison project [CMIP] simulations [Eyring *et al.*, 2016]), which captures uncertainty regarding the model structure and parameterisation assumptions (structural uncertainty [Rougier, 2007]). Within a given model, there are also uncertainties around the model parameters that define parameterisations, which we call parametric uncertainty. These can be quantified by running 'Perturbed Parameter Ensembles' (or 'Perturbed Physics Ensembles', PPEs) which involve sampling model parameters according to domain expertise and running ensembles of simulations (e.g., Christensen *et al.*, 2015b; Eidhammer *et al.*, 2024; Murphy *et al.*, 2004, 2007; Sengupta *et al.*, 2021; Stainforth *et al.*, 2005).

1.3 | Reducing uncertainties in general circulation models

While uncertainties can never be fully eliminated, parametric uncertainties can be reduced in a process known as calibration, a key aspect of UQ and model development of both weather and climate models (e.g., Carslaw *et al.*, 2013; Dunbar *et al.*, 2021; Sengupta *et al.*, 2021; Souza *et al.*, 2020; Williamson *et al.*, 2017). Calibration focuses on reducing parametric uncertainty, by constraining parameters based on past observations. Techniques that leverage PPEs are often used to eliminate parameter values that produce model output inconsistent with observations, through history matching (Couvreux *et al.*, 2021; King *et al.*, 2024; Raoult *et al.*, 2024; Williamson *et al.*, 2013), approximate Bayesian computation (Watson-Parris *et al.*, 2021), or ensemble Kalman methods (Dunbar *et al.*, 2021; Mansfield & Sheshadri, 2022). Many of these methods are aided by machine-learning emulators such as Gaussian process emulators, which reduce the number of expensive GCM integrations required. In contrast to this, initial-condition uncertainty, subgrid variability and internal variability are internal properties of the earth system and our observations of it and therefore cannot be reduced through model development.

1.4 | The rise in machine-learning parameterisations

Over the last few years, we have witnessed a rise in ML-based parameterisations, which are trained on existing physics-based parameterisations (e.g., Chantry *et al.*, 2021; Espinosa *et al.*, 2022; Ukkonen, 2022), cloud-resolving models embedded within GCM grid cells (e.g., Hu *et al.*, 2025; Rasp *et al.*, 2018; Yu *et al.*, 2024), coarse-grained high-resolution GCM simulations (e.g., Giles *et al.*, 2024; Grundner *et al.*, 2022; Henn *et al.*, 2024; Heuer *et al.*, 2024; Morcrette *et al.*, 2025; Ross *et al.*, 2023; Watt-Meyer *et al.*, 2024; Yuval & O'Gorman, 2023), or observations (e.g, Miller *et al.*, 2025). There has been substantial interest in stochastic ML-based parameterisations, which includes stochasticity in a similar manner to described above (Christensen *et al.*, 2024). Stochastic ML schemes are primarily used to improve model skill (e.g., Gagne II *et al.*, 2020; Giles *et al.*, 2024; Guillaumin & Zanna, 2021; Nadiga *et al.*, 2022; Perezhogin *et al.*, 2023) but some studies also focus on their use for UQ (e.g., Behrens *et al.*, 2025; Mansfield & Sheshadri, 2022; Miller *et al.*, 2025). Here, we argue that while AI is becoming increasingly used in weather and climate modelling, evaluating and quantifying uncertainty remains critical for ensuring trust and credibility in predictions (Haynes *et al.*, 2023; McGovern *et al.*, 2022). As we demonstrate here, probabilistic ML methods, such as Bayesian deep learning, also provide a natural framework for UQ.

1.5 | This study

In this study, we quantify uncertainties in a subgrid parameterisation, decomposing them by source and across different time-scales. We use the Lorenz, 1996 (L96) model as a case study to demonstrate how UQ should be carried out and to compare different approaches to sampling uncertainties. We use L96 because it simulates both large- and small-scale variables and their interactions and exhibits chaotic properties similar to that of the real atmosphere but can be run at a significantly lower computational cost than a full GCM. This makes it a suitable testbed for parameterisations. It has been used to test stochastic parameterisations (Arnold *et al.*, 2013; Wilks, 2005) and machine-learning parameterisations (Chattopadhyay *et al.*, 2020; Gagne II *et al.*, 2020; Parthipan *et al.*, 2023; Rasp, 2020) for numerical weather prediction. It has also been used to explore the uncertainties associated with parameterisations on climate time-scales, for instance, in PPE studies (Christensen *et al.*, 2015a). Although a simplified model, we can use it to draw analogies to more complex GCMs, while considering a range of prediction time-scales.

In Section 2, we discuss the framework of uncertainties commonly used in the ML community that describe aleatoric uncertainty, coming from the data, and epistemic uncertainty coming from the model, and how these fit in with the traditional uncertainty viewpoint used in the weather and climate communities. In Section 3, we outline model and methods used, including the Lorenz 1996 model and how a deterministic neural-network parameterisation can be used in place and how we quantify uncertainties using Bayesian neural networks (BNNs), including an assessment of uncertainties in an 'offline' setting (i.e., data pre-generated). In Section 4, we analyse the uncertainties once the parameterisations are coupled back in the Lorenz 1996 model, which we refer to as 'online'. We consider these on weather forecasting time-scales and in Section 5, we consider these on a climate time-scale. Finally, Section 6 draws conclusions and discusses how we can assess uncertainties in weather and climate models that use ML parameterisations going forward.

2 | BACKGROUND

2.1 | Types of uncertainties in machine learning

Here, we approach UQ from a machine-learning perspective, where we categorise uncertainty into two types: aleatoric uncertainty and epistemic uncertainty (Hüllermeier & Waegeman, 2021).

The term 'aleatoric' comes from the Latin word *alea*, meaning 'game of chance'. Aleatoric uncertainty is used to describe the variability in a system that is due to inherently random effects (Haynes *et al.*, 2023; Hüllermeier & Waegeman, 2021). It represents the statistical or stochastic nature of a system, such as flipping a coin or rolling a dice and this type of uncertainty cannot be reduced. In the ML literature, aleatoric uncertainty refers to uncertainty in the data. This could include data generated from a stochastic process, but it can also include data originating from unobserved variables, which means there is no longer a one-to-one mapping from inputs to outputs.

The term 'epistemic' comes from the Greek word *epistēmē*, which means knowledge or understanding. Epistemic uncertainty is caused by a lack of knowledge about the best model for a system. This can be thought of as a systematic uncertainty and can be reduced with more knowledge or understanding of the system which can come from more data. In the ML literature, epistemic uncertainty refers to uncertainty in the ML model. This includes uncertainty in model structure (i.e., model architecture and number of layers and neurons), and uncertainties in model parameters (i.e., weights and biases).

2.2 | Aleatoric and epistemic uncertainties in general circulation models

We can broadly relate these types of uncertainties to those used by the earth system modelling community. Initial-condition uncertainty and internal variability can be viewed as both forms of aleatoric uncertainty, as they arise from the chaotic nature of the earth system. Here, we will also treat subgrid variability as aleatoric uncertainty because it arises from the internal variability that can occur within a grid box. From an ML perspective, this can be interpreted as uncertainty in the training data that arises when knowledge of the large-scale state, X, does not uniquely define the subgrid-tendency, U. This type of uncertainty is irreducible given the task is to learn the U from X alone. In contrast, we will consider structural and parametric uncertainties as epistemic uncertainty, because they are modelling choices made during model development. We also consider scenario uncertainty predominantly as a form of epistemic uncertainty, as it comes from lack of knowledge in the appropriate external forcing. We emphasise that this distinction is not sharp, and many quantities may contain both aleatoric and epistemic components, for instance, initial-condition uncertainty arises from first, lack of knowledge of the initial conditions, combined with the chaotic nature of the atmosphere and sensitivity to these conditions. These classifications help organise our thinking about uncertainty, rather than

TABLE 1 Types of uncertainties in GCMs on varying time-scales, whether they are aleatoric or epistemic, and the traditional approach used in weather and climate.

Time-scale	Type of uncertainty in GCMs	Aleatoric	Epistemic	Typical approach
Weather	Initial condition	x		Perturbed IC ensemble (atmosphere, derived from observations)
Weather	Subgrid variability (informs parameterisations, i.e., training data)	x		Stochastic parameterisation
Seasonal to decadal	Internal variability	x		Perturbed IC ensemble (for longer time-scales, perturbed sea surface temperatures)
All	Structural uncertainty		x	Multimodel ensemble
Climate	Parametric uncertainty		x	Perturbed-parameter ensemble
Climate	Scenario uncertainty		x	Multiscenario ensemble

Abbreviations: GCMs, general circulation models; IC, initial condition.

act as strict divisions. These types of uncertainties, the time-scales they generally dominate on, and the typical approach used to estimate them are summarised in Table 1.

2.3 | Aleatoric and epistemic uncertainties in general circulation models that use machine-learning parameterisations

Machine learning is becoming increasingly used for parameterisations in GCMs. We expect ML-based parameterisations to have similar associated uncertainties to physics-based schemes. In the ML framework, the goal is to learn the relationship between the large-scale variables and the subgrid variables from the training data. The subgrid variability creates a noisy dataset and can be viewed as the uncertainty in the training data, in other words, the aleatoric uncertainty.

Epistemic uncertainties also exist within ML-based parameterisations that are analogous to the types of uncertainties discussed previously for conventional parameterisations. While physics-based parameterisations typically only contain a handful of parameters that lead to parametric uncertainty, ML-based parameterisations have a large number of parameters to learn. This means the parametric uncertainties and the relationships between them are likely more complex. Structural uncertainties are also present for ML-based parameterisations, since there are choices to make regarding ML algorithms, architectures, and number of parameters.

Furthermore, the use of ML gives rise to another form of epistemic uncertainty that isn't traditionally considered in UQ: '*out-of-regime*' or '*out-of-sample*' uncertainty. This occurs when an ML algorithm is trained on data that is generated by one distribution but is applied to a dataset that was generated by a different distribution. ML algorithms are known to perform poorly during extrapolation, especially if they have a large number of parameters and/or have been overfit. This could be an issue for climate modelling, where training data may come from present-day climate, but the ML parameterisation could be applied under a different future climate scenario.

In this study, we quantify epistemic and aleatoric uncertainties in an ML parameterisation. We use BNNs to capture both epistemic uncertainty, which represents parametric uncertainty and out-of-regime uncertainty, and aleatoric uncertainty, which represents the subgrid variability. BNNs provides us a with a way to clearly distinguish between these two forms of uncertainty. Note that within epistemic uncertainty, we consider only parameteric uncertainty given a fixed neural-network architecture. This means we do not consider *structural uncertainty* which arises from different modelling choices, or *model discrepancy* which arises from the fact that the neural network is a surrogate model that may not be able to perfectly represent the true underlying dynamics. Within aleatoric uncertainty, we assume that the training data are fixed and do not consider data biases that could exist within this limited dataset.

3 | METHODS

3.1 | The Lorenz 1996 model

We adopt the two-layer L96 system as a simplified model of chaotic dynamics to explore UQ methods for ML parameterisations (Lorenz, 2006). It can be viewed as toy model

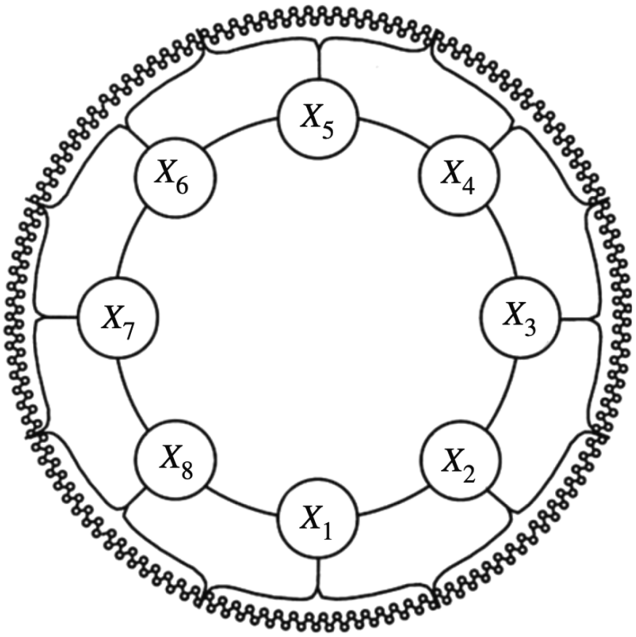

FIGURE 1 Schematic showing the two-layer L96 setup with $K = 8$ large-scale variables, X, and $J = 32$ small-scale variables, Y, coupled to each X variable and cyclic boundary conditions. Reproduced from Wilks (2005).

for mid-latitude atmospheric dynamics around a latitude circle. In the one-layer version, the variables are X_k where $k = 1, \ldots, K$, with periodic boundary conditions, that is, $X_{K+1} = X_1$. The variables evolve following

$$\frac{dX_k}{dt} = -X_{k-1}(X_{k-2} - X_{k+1}) - X_k + F, \tag{1}$$

where F is an external forcing. This can be solved with a choice of numerical methods, such as the Euler method.

The two-layer version extends this to include large-scale variables, X, and small-scale variables, Y, defined from $j = 1, \ldots J$, between each of the large-scale variables (Figure 1) (Wilks, 2006).

$$\frac{dX_k}{dt} = -X_{k-1}(X_{k-2} - X_{k+1}) - X_k + F + \frac{-hc}{b}\sum_{j=J(k-1)+1}^{kJ} Y_j, \tag{2}$$

$$\frac{dY_j}{dt} = -cbY_{j+1}\left(Y_{j+2} - Y_{j-1}\right) - cY_j + \frac{hc}{b}X_{\text{int}[j-1/J]+1}, \tag{3}$$

where the int. notation refers to the integer value of the term inside the brackets (i.e., the closest X variable) and b, c, and h are user-defined parameters of the system that describe the spatial scale ratio, the temporal scale ratio and a coupling constant, respectively. Following Arnold *et al.* (2013) and Wilks (2005), we use $K = 8, J = 32, b = 10$, $c = 10, h = 1$ and $F = 20$. The term highlighted in red is the subgrid-scale term that couples the small-scale variables to the large-scale variables.

Since the two-layer system must solve for the small-scale variables, it usually requires more advanced numerical methods such as fourth-order Runge–Kutta Scheme (RK4), used here with a timestep of $\Delta t = 0.001$. This motivates the task of replacing the small-scale variables with a parameterisation, allowing a coarser timestep and cheaper numerical scheme to be used.

To reduce computational cost, we can replace the coupled system of equations with a single equation that describes only the large-scale variables with an additional parameterisation term, $U_k = f(X_k)$, as follows:

$$\frac{dX_k}{dt} = -X_{k-1}(X_{k-2} - X_{k+1}) - X_k + F + f(X_k). \tag{4}$$

This is the 'forecast model'. This parameterisation term aims to capture the small-scale variables, Y, without explicitly evaluating Equation (2). The function f can be learned from a training dataset. This can then be solved with a cheaper numerical method such as RK2, and with a coarser forecast timestep, $\Delta t_f = 0.005 = 5\Delta t$, reducing computational burden by around five times, assuming the function f is cheap to evaluate.

3.2 | Training a neural-network parameterisation

To learn the function, f, we can train a neural network. Training data can be generated by simulating the two-layer coupled system and to use these X_k terms to learn $U_k = f(X_k)$ in the forecast model (Equation 4). Rather than directly storing the subgrid-scale term ($\frac{-hc}{b}\sum_{j=J(k-1)+1}^{kJ} Y_j$ in Equation 1), as done in Balwada *et al.* (2024) and Rasp (2020), we estimate the subgrid-scale term from the X_k, as done in Arnold *et al.* (2013), Gagne II *et al.* (2020), Parthipan *et al.* (2023), and Wilks (2005). Given a dataset X_k from the two-layer model, stored at intervals Δt_f, we can rearrange Equation (4) to obtain.

$$[U_k]_t = \frac{[X_k]_{t+\Delta t_f} - [X_k]_t}{\Delta t_f} - [-X_{k-1}(X_{k-2} - X_{k+1}) - X_k + F]_t. \tag{5}$$

This is representative of how parameterisations are trained on high-resolution datasets, where the subgrid tendencies are not available and instead must be estimated through carrying out a coarse-graining in space (e.g., Heuer *et al.*, 2024; Morcrette *et al.*, 2025; Watt-Meyer *et al.*, 2024;

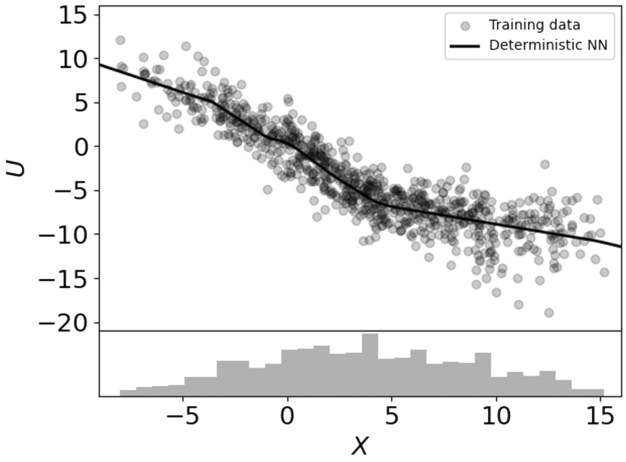

FIGURE 2 Training data where the X-axis shows the large-scale variables (inputs to neural network) and the Y-axis shows the subgrid tendency (outputs of neural network). The black solid line shows the prediction from the deterministic neural network. The histogram shows the distribution of the data.

Yuval & O'Gorman, 2023). ML parameterisations typically use only variables from a single vertical column, since information from neighbouring grid cells is unavailable. Consistent with this approach, our goal is to predict U_k as a function of X_k alone. We will treat all K points on the circle as separate data points. For the training dataset, we select 100 independent data points from a time series over a span of $T = 1000$ model time units (MTU), equivalent to around 1½ 'atmospheric years', estimated by considering the error-doubling time in the system (Wilks, 2005). Including the spatial dimension with $K = 8$, this gives 800 samples for training, shown in Figure 2. Other ML parameterisation studies typically use $O(10^6 - 10^7)$ samples to train networks with $O(10^5 - 10^6)$ parameters (e.g., Grundner *et al.*, 2022; Ukkonen & Chantry, 2025; Yu *et al.*, 2024), giving comparable sample-to-parameter ratios.

We start with a basic deterministic fully connected neural network with two hidden layers, each with 16 nodes and ReLU activation functions, (the same architecture as Rasp, 2020, but with 16 rather than 32 nodes per layer). A small network is suitable given past studies show a linear regression or cubic polynomials also perform well on this dataset (e.g., Arnold *et al.*, 2013; Parthipan *et al.*, 2023; Wilks, 2006). We write a neural network as: $U = f_\theta(X)$, where X are the inputs, U are the outputs and θ are the network weights. We train the network to minimise mean squared error. The black line in Figure 2 shows the resulting neural-network fit. This neural network does pick up the general trend, but it is evident that there is noise within the data coming from the subgrid variability that cannot be captured by the deterministic neural network. In the following section, we train a BNN to capture this as a form of aleatoric uncertainty, as well as quantifying epistemic uncertainties from uncertainty in the model parameters. We use the same architecture as the deterministic neural network.

3.3 | Learning uncertainties using a Bayesian neural network

To quantify and separate the sources of uncertainty into epistemic and aleatoric, we use BNNs. BNNs extend traditional neural networks with a Bayesian treatment of the neural-network parameters (the weights and biases, θ). We use this approach because in recent years, neural networks have become the primary choice of machine-learning parameterisations (e.g., Hu *et al.*, 2025; Ukkonen & Chantry, 2025; Yuval & O'Gorman, 2023), due to their flexibility, simplicity to implement, and widespread use in the machine-learning community. We will write a Bayesian neural network (BNN) as: $U = f(X|\theta)$.

Instead of learning fixed weights, BNNs treat the weights as probability distributions and use Bayes' theorem to update these distributions based on observed data (Goan & Fookes, 2020). This captures uncertainty in the parameters or *parametric uncertainty*, a form of epistemic uncertainty. Since this approach is probabilistic, the approach to training differs from standard neural-network training. Rather than seeking *parameters* that best fit the data, θ, we seek *probability distributions*, $p(\theta|X, U)$ that are most likely to generate the dataset (X, U). To do this, we must define prior distributions on θ and carry out Bayesian inference to learn the posterior probability distribution. Here, we specified a zero-mean isotropic Gaussian prior for each θ, and used variational inference to learn the posterior distribution, which provides a computationally efficient approach for sampling (Blei *et al.*, 2017; Ranganath *et al.*, 2013). This approximates the posterior with a variational distribution, here a multivariate Gaussian with a full covariance structure across all network parameters (Barber & Bishop, 1997; Hinton & van Camp, 1993). See Supplementary Text S1 for a full description of variational inference.

Once we have learned the parameter probability distribution, we can evaluate our BNN at new data points, X^*, to obtain the posterior predictive distribution,

$$p\left[f(X^*)\right] = \int p\left[\left[f(X^*|\theta)\right]\, p(\theta|X, U)\, d\theta, \qquad (6)$$

where we integrate over all the possible values for θ.

Figure 3 shows a simplified schematic of the BNNs used here. The distributions on the connections between nodes highlight that all parameters (weights and biases)

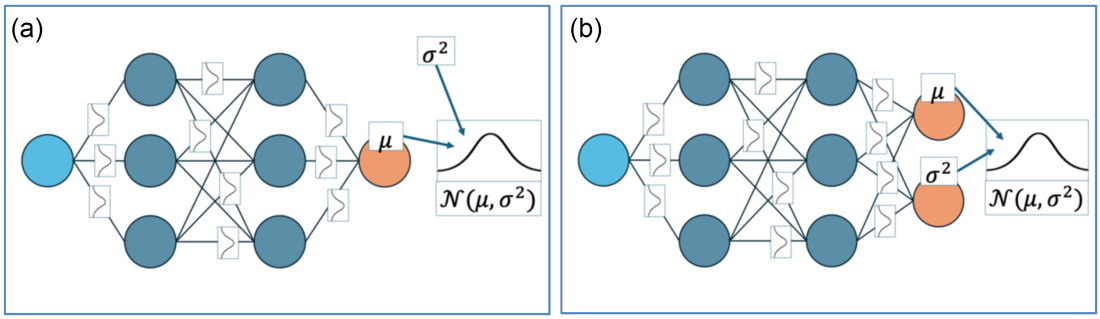

FIGURE 3 Schematic of Bayesian neural network. All network weights are assumed to have associated probability distributions, represented by the normal distributions on the nodes in this diagram (epistemic uncertainty). The output layer is also assumed to be a normal distribution with mean μ and variance σ^2 which represents the noise in the data (aleatoric uncertainty). The BNN in (a) learns σ^2 as a fixed scalar, independent of input values (known as 'homoscedasticity') while in (b) the BNN directly outputs σ^2, allowing it to depend on the inputs (known as 'heteroscedasticity'). Note that this is a simplified representation of the BNN used here, as we use more nodes within each hidden layer and the weights are treated as multivariate normal distributions to capture correlations between the weights.

have associated probability distributions. This directly captures *parametric uncertainty*. We can also treat the model output as a probability distribution, allowing us to capture inherent noise in the dataset, or *aleatoric uncertainty*. This means the BNN, $f(X|\theta)$, predicts not just a single value, but a probability distribution representing the dataset. Here, we represent the data with a Gaussian distribution, with mean μ and the variance σ^2. Figure 3a shows that the output layer predicts this mean μ and we separately learn σ^2 as a fixed parameter across the entire dataset. This does not consider that some regions of the input space may be associated with increased subgrid variability, which should be represented as higher aleatoric uncertainty. Figure 2 above shows that this may indeed be case, as there is increased noise for larger values of X. Allowing the aleatoric uncertainty to vary with X is known as heteroscedasticity, and is a common challenge in Bayesian machine learning (Kendall & Gal, 2017). We will consider a heteroscedastic version of the BNN by learning the variance on the output layer as a function of the inputs, shown in Figure 3b. The BNN uses the same architecture but now predicts two values as outputs instead of just one: the mean μ and the variance σ^2. We refer to this the heteroscedastic BNN. This approach is more complicated because the aleatoric uncertainty now depends upon the parametric uncertainty within the BNN, meaning there is not such a clear separation between the two forms of uncertainty.

3.4 | Sampling epistemic and aleatoric uncertainties offline

Once the BNN is trained, we can sample from the parameter distributions to estimate epistemic and aleatoric uncertainties. For a given value of X, we must sample the parameters θ from the posterior distribution and evaluate the neural network for each θ, to obtain $(U|X, \theta)$. Then we can use the law of total variance to decompose uncertainty into epistemic and aleatoric components (Valdenegro-Toro & Mori, 2022):

$$\mathrm{Var}(U|X) = \underbrace{E_\theta[\mathrm{Var}(U|X,\theta)]}_{\text{Aleatoric}} + \underbrace{\mathrm{Var}_\theta(E[U|X,\theta])}_{\text{Epistemic}}. \quad (7)$$

To sample *epistemic variance*, we compute the variance across the mean component of the output layer, $(\mu \mid X, \theta)$. Alternatively, we can sample *aleatoric variance* by computing the mean over the variance component of the output layer $(\sigma^2 \mid X, \theta)$.

We first assess the uncertainties in BNN predictions in an *offline* setting, meaning the input X data has already been generated by the full two-layer system in Equations (1) and (2) (Bracco *et al.*, 2025). Figure 4 shows the BNN mean prediction across the domain and the shading shows two standard deviations uncertainty obtained when sampling from aleatoric, epistemic or both sources of uncertainty as described above. The points show the training data (same as Figure 2). Figure 4a shows the results of homoscedastic BNN that assumes fixed aleatoric uncertainty across the entire dataset (Figure 3a), while Figure 4b shows the heteroscedastic version where the BNN also predicts the variance as a function of the input data, allowing the aleatoric uncertainty to vary with X (Figure 3b). The latter appears to better capture the variations in the data. Both show epistemic uncertainty to be lower in the centre of the dataset where the parameters are more constrained but increase as towards the edges of the dataset where there is increased out-of-regime uncertainty, although overall, epistemic uncertainty only has a small contribution to the total uncertainty. For the

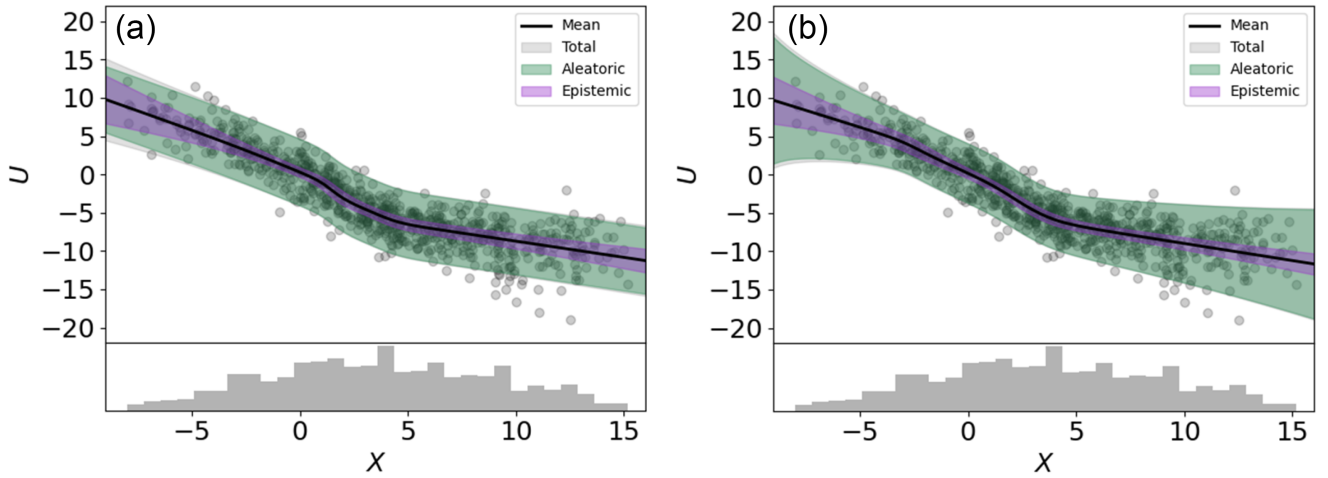

FIGURE 4 Offline results for aleatoric, epistemic, and total uncertainty, where shading shows two standard deviations away from mean. (a) Aleatoric uncertainty is learned as a scalar parameter that is fixed across the entire dataset (homoscedastic). (b) Aleatoric uncertainty varies across the dataset (heteroscedastic) by setting the Bayesian neural network (BNN) up to predict two outputs: a mean and a variance. The grey histogram shows the distribution of the data.

rest of the paper, we use heteroscedastic BNN because it appears to capture aleatoric uncertainty better, unless stated otherwise.

We explored different training approaches and found the BNN to be robust against different choices of priors (Figure S1). Increasing the size of the training dataset also did not significantly reduce epistemic uncertainty (Figure S2). For the variational distribution, we found that a mean-field approximation (i.e., treating all parameters as independent) would not be sufficient as there are significant covariances between neural-network parameters (Figure S3). We also compared variational inference against direct Bayesian inference using Monte Carlo Markov chain, where we found similar total predictive variance but variational inference estimates higher epistemic uncertainty (Figure S4). This may reflect approximation error from assuming that the variational family (here, a multivariate Gaussian) can approximate the true posterior distribution, While still a form of epistemic uncertainty, this reflects uncertainty in the methodology rather than parametric uncertainty (Valdenegro-Toro & Mori, 2022). Although we acknowledge this additional uncertainty, we continue to use variational inference for pragmatic reasons, as it simplifies sampling during online simulations which would make it more suited for use in full GCM simulations.

4 | ONLINE COUPLING ON WEATHER TIME-SCALES

After training, we couple the machine-learning parameterisation back into the one-layer L96 through Equation (3) where $f(X)$ is the trained ML algorithm. This is used to update the state X at the next timestep, which in turn is used to estimate $f(X)$ at the following timestep. This creates a feedback between the large-scale dynamics and ML prediction of the subgrid-scale physics. We call this *online* evaluation (Bracco *et al.*, 2025).

4.1 | Sampling epistemic and aleatoric uncertainties online

Ultimately, we are interested in how epistemic and aleatoric uncertainties influence the output of the coupled dynamical system, $X_k(t)$, rather than only the parameterisation output U. To quantify this, we use the BNN as a stochastic parameterisation and generate ensemble members under three configurations:

i. Total uncertainty: at each timestep, we draw θ from the posterior distribution; evaluate the BNN to obtain $\left[\mu, \sigma^2 | X, \theta\right]$; then sample $U \sim N\left(\mu, \sigma^2\right)$.
ii. Epistemic uncertainty: we draw θ from the posterior distribution; evaluate the BNN, but only use the mean prediction μ, ignoring the variance term σ^2.
iii. Aleatoric uncertainty: we fix θ at the posterior mean values, $\overline{\theta}$ (obtained from the variational distribution, see Supplementary Text S2 for details); then evaluate the BNN to estimate $\left[\mu, \sigma^2 | X, \overline{\theta}\right]$, then sample $U \sim N\left(\mu, \sigma^2\right)$.

We use ensemble spread in the three configurations to measure total, epistemic and aleatoric uncertainty, respectively. Note that these differ slightly from the offline

decomposition. For epistemic uncertainty, we use just one sample at each timestep, but over the duration of the simulation we expect to sample the full posterior so that ensemble spread in the trajectory represents $\mathrm{Var}_\theta(E[Y|X,\theta])$ in Equation (7). This is representative of Stochastically Perturbed Parametrisations (SPP) which treat uncertain parameters as a stochastic process and sample them throughout a forecast to capture model uncertainty (Lang *et al.*, 2021). To sample aleatoric uncertainty, we do not directly estimate estimate $E_\theta[\mathrm{Var}(U|X,\theta)]$ in Equation (7). but instead use $Var(U|\mathrm{X},\overline{\theta})$ to represent aleatoric variance, where $\overline{\theta}$ is fixed at the mean values throughout the duration of all simulations. This is done for consistency with both operational weather parameterisations, such as SPPT (Buizza *et al.*, 1999) and with previous stochastic ML-parameterisations studies (e.g., Guillaumin & Zanna, 2021; Zhang *et al.*, 2023). The philosophy of these frameworks considers parameters to be deterministic, with appropriate perturbations added to capture subgrid variability. It is also a pragmatic approach because repeatedly sampling θ would be computationally prohibitive in an operational weather or climate model. Figure S5 confirms that $\left[\sigma^2|X,\overline{\theta}\right]$ is generally a good approximation of the true aleatoric variance, $E_\theta\left[\sigma^2|X,\theta\right]$ across most of the input space; however, towards the edges of the input space when $X<-5$, $\left[\sigma^2|X,\overline{\theta}\right]$ underestimates $E_\theta\left[\sigma^2|X,\theta\right]$. This is because some parameter settings predict a very high variance, given the limited training data in this regime, indicating that larger epistemic uncertainty can lead to larger aleatoric uncertainty. This highlights the challenges of cleaning separating these sources of uncertainty. We note this as a limitation of our approach, although in practice, in this L96 setup, X rarely drops below 5 (less than 4% of the training data), so this should not have a major effect on our results.

For the weather forecasting problem, we use the BNN to forecast trajectories that sample epistemic, aleatoric and total uncertainty. We compare these to a 'truth' experiment that uses the two-layer L96 model, initialised with the state at the end of the training dataset and run for $T=1000$. We identify $N_{\mathrm{init}}=100$ initial conditions from the truth dataset, separated by $T=10$ (which corresponds to about 50 atmospheric days, more than sufficient for the initial conditions to be uncorrelated), and generate $N_{\mathrm{ens}}=50$ ensemble members for each configuration.

4.2 | Independent noise at each time step

At each time step, to estimate U in Equation (3) with the BNN, $U=f(X|\theta)$ we must sample the weights θ. This creates a stochastic parameterisation where we obtain a different outcome each time we estimate U for a given X. To estimate uncertainties, we must run multiple ensemble members, here using $N=50$. We run the BNN in three configurations: 'aleatoric', where the model parameters are kept fixed at their median values but we sample from the output layer ($\mathcal{N}\left(\mu,\sigma^2\right)$ in Figure 3b), 'epistemic', where the model parameters are sampled and the output layer is deterministic (μ in Figure 3b), and 'both', where we sample from the model parameters and the output layer. Here, the sampling methods are entirely independent at each time step (i.e., a white noise process). This means that, unlike most stochastic parameterisations (e.g., SPPT, Buizza *et al.*, 1999), the random component of the subgrid-scale prediction is not correlated in time.

Figure 5a,c,e shows the trajectories of one variable given one initial condition under each of these settings. We find that the ensemble members start to diverge from each other after $T\sim0.6$. They diverge from the truth slightly earlier, after $T\sim0.4$, indicating poor skill. The epistemic uncertainty is significantly lower than the aleatoric uncertainty and takes longer for the ensemble members to diverge. 'Both' appears to follow the aleatoric uncertainty more closely.

4.3 | Including temporal correlation

Figure 5a,c,e show the ensemble members diverge from the truth before they spread out from each other. Unlike many stochastic parameterisations used in operational weather forecasting models (Buizza *et al.*, 1999; Shutts, 2005), there is no temporal correlation between U_k. Many studies have found that temporal correlation is essential in stochastic parameterisations to both improve skill and the spread of a forecast (Arnold *et al.*, 2013; Berner *et al.*, 2017). In Figure 5b,d,f, we include temporal correlation following an Auto Regressive Order 1 (AR1) process, similar to the method used in Arnold *et al.* (2013).

The AR1 approach assumes that the stochastic parameterisation can be broken down into two parts: a deterministic component (here, the neural network with weights fixed at their median values) and a random-noise component (Wilks, 2006). To maintain correlation with the previous time step, the random component combines the noise from the previous time step with a new noise term, known as the *innovation*. We use the autoregressive parameter, estimated from the lag-1 correlation (0.985) to enforce a suitable correlation between successive time steps, giving a correlation time-scale of 0.33 MTU. To define the innovation to be added at each time step, we estimate the variance associated with the aleatoric or epistemic uncertainty for the input, X_t. Full details are provided in the Supplementary Text S2.

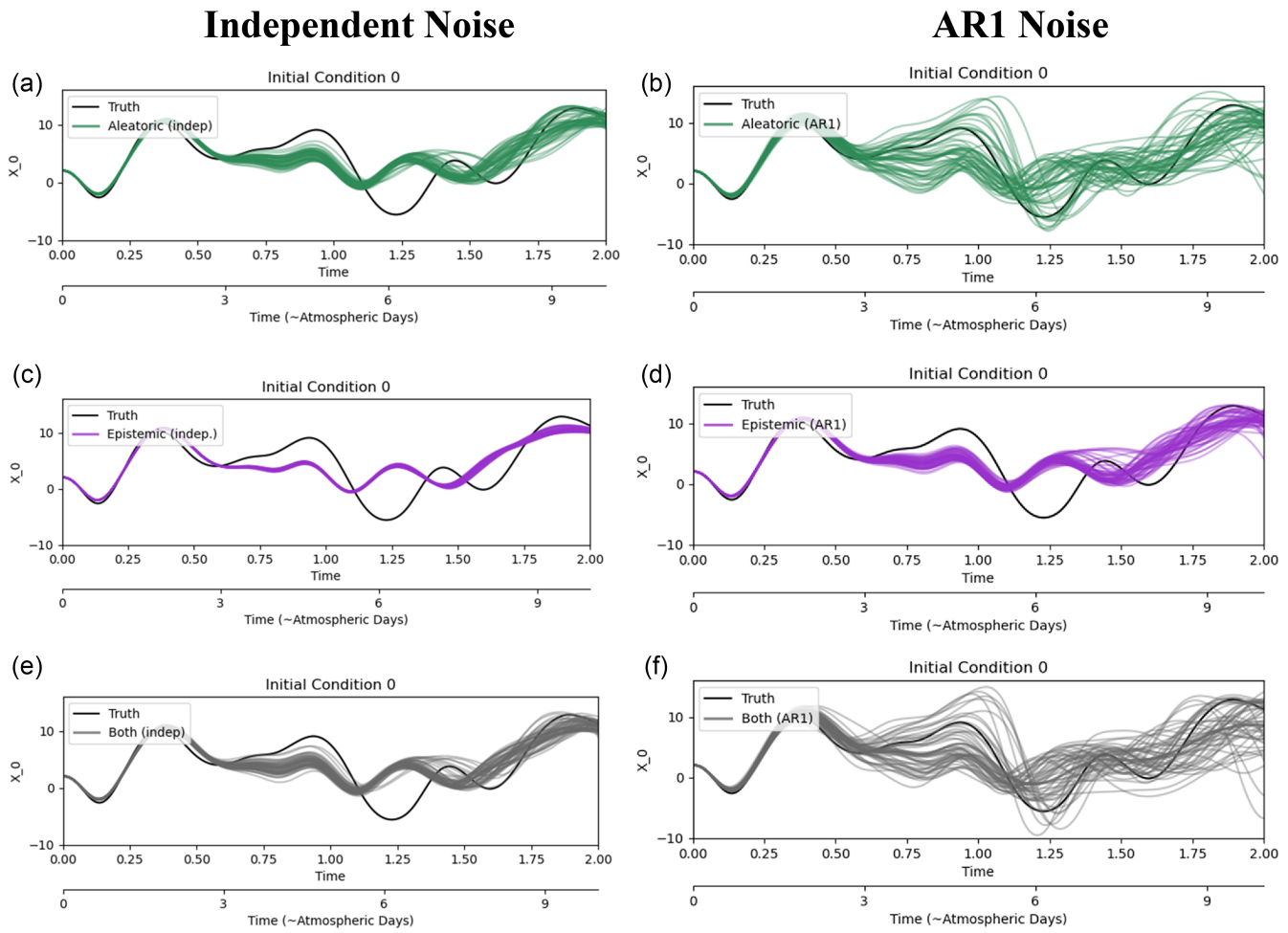

FIGURE 5 Trajectories for variable X_0, generated by the one-layer L96 model with the stochastic parameterisation where each line represents a different ensemble member and the black line shows the true, two-layer L96 model. The parameterisations sample (a,b) aleatoric uncertainty only, (c,d) epistemic uncertainty only and (e,f) both types of uncertainty. Panels on the left (a,c,e) show parameterisations that are sampled independently at each time step, while panels on the right (b,d,f) use an Auto Regressive Order 1 (AR1) process to include temporal correlations.

Figure 5b,d,f show the trajectories that use the AR1 parameterisation for aleatoric, epistemic, and both sources of uncertainty, respectively. As before, simulations that sample both types of uncertainty have similar spread to those that sample aleatoric uncertainty only. For all simulations, using an AR1 sampling approach leads ensemble members to diverge faster than with independent sampling. The ensemble members spread out enough to capture the truth over much of the simulation, showing a major improvement in the reliability of the forecast. Note that we see similar results when using the homoscedastic BNN, although adding the AR1 process increases aleatoric uncertainty more significantly (Figure S6).

We test how well the ensemble members agree with the truth more robustly by repeating simulations with multiple different initial conditions. Figure 6 shows the overall RMSE and spread against time in solid and dashed lines respectively. These are computed as the square root of the sum of the squared error (RMSE) and variances (spread) over 100 simulations with different initial conditions. Figure 6a shows that when we sample independent noise at each time step, the spread is consistently smaller than the RMSE, indicating underdispersive ensembles. When AR1 noise is introduced, Figure 6b shows that the spread and error match well and grow at similar rates. The errors also grow more slowly. Note this is not the case for the homoscedastic ensembles, where including aleatoric uncertainty leads to an overdispersive ensemble (Figure S7). In all cases, the spread in the aleatoric ensembles is almost identical to the spread in the ensembles that sample both forms of uncertainty. This suggests that they are not distinct, independent sources of uncertainty and that the contribution from epistemic uncertainty is small enough to ignore on these time-scales.

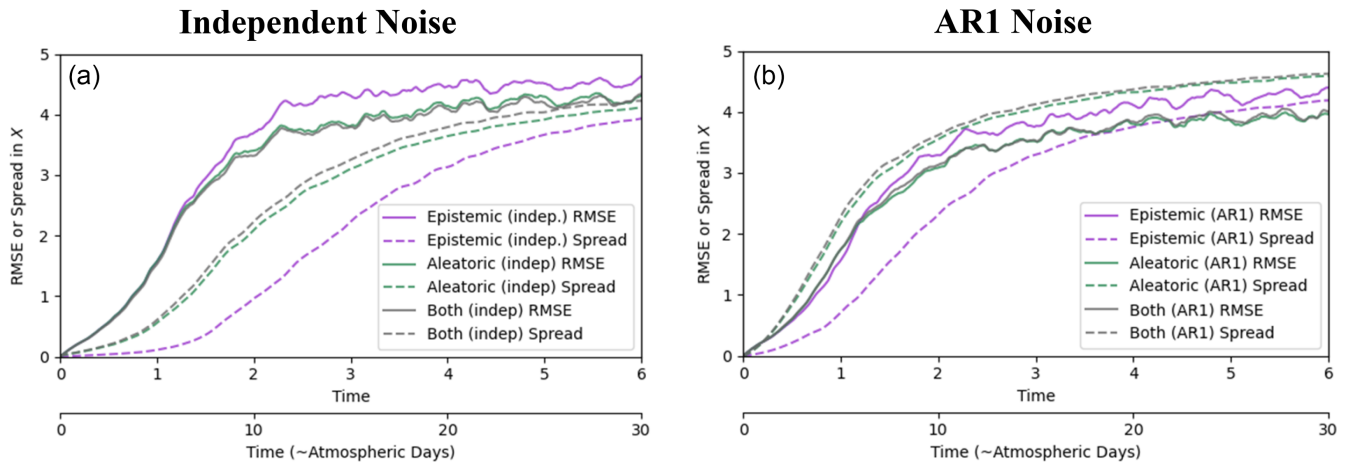

FIGURE 6 Root of the sum of the squared error (RMSE) (solid line) and spread (dashed line) both averaged over 100 different simulations for (a) independent noise and (b) AR1 noise.

This confirms the validity of past studies on ML stochastic parameterisations for weather time-scales, where only aleatoric uncertainty from the subgrid variability is captured by the ML scheme (Behrens *et al.*, 2025; Guillaumin & Zanna, 2021; Perezhogin *et al.*, 2023).

4.4 | Reliability

The reliability of a forecast refers to how well predicted probabilities align with observed probabilities (Arnold *et al.*, 2013; Leutbecher, 2009; Leutbecher & Palmer, 2008). For ensemble forecasts, the ensemble spread should be an indication of the error in the ensemble mean. This is known as statistical consistency. We test for statistical consistency by considering spread against error for the 100 independent forecasts, evaluated at the same time ($t = 0.5$ MTU, approximately $2\frac{1}{2}$ atmospheric days). Following Leutbecher (2009), we compute the ensemble mean and variance for each sample. We sort the samples by increasing variance and partition them into bins of size 100. Root mean squared (r.m.s.) spread is defined as the square root of the mean ensemble variance within each bin and r.m.s. error is defined as the square root of the variance of the ensemble mean error within each bin. Figure 7 shows the r.m.s. spread on the X-axis against r.m.s. error on the Y-axis for (a) the independent noise simulations and (b) the AR1 simulations. Points lying in the upper left triangle of the plot indicate underdispersive ensembles, where spread is less than error, whereas the points lying in the lower right triangle are a feature of an overdispersive ensemble, where spread is larger than error. Points lying along the $y = x$ line indicate a well-calibrated ensemble. Figure 7a shows all independent noise simulations fall within the upper left triangle, indicating underdispersive ensembles, while Figure 7b shows that using the AR1 process increases the spread significantly. Although the epistemic ensembles are still underdispersive, when aleatoric uncertainty is included, the ensembles appear better calibrated, falling closely along the $y = x$ line. Figure 7c,d shows a similar story for the homoscedastic BNN but the constant aleatoric uncertainty leads to slightly overdispersive ensembles when including the AR1 process. This suggests that a heteroscedastic treatment of aleatoric uncertainty is important for well-calibrated ensembles.

5 | ONLINE COUPLING ON CLIMATE TIME-SCALES

5.1 | Climate simulations

For climate prediction, we are interested in long-term statistics, so for this we run a long simulation with $T = 1000$, corresponding to around 14 years. Figure 8a shows that over these time-scales, the distributions of X for all simulations match the truth closely but they all slightly underestimate the tails of the distribution, highlighted by the lower panel which shows the difference between the distribution and the true distributions. There are no major differences between any sampling approaches. The deterministic parameterisation performs well and behaves very similarly to the epistemic parameterisation, likely because the epistemic uncertainty introduced is small (especially when using the AR1 process, which has a short correlation time-scale [~0.33 MTU] and therefore its effect averages out over longer time-scales). This highlights the need for

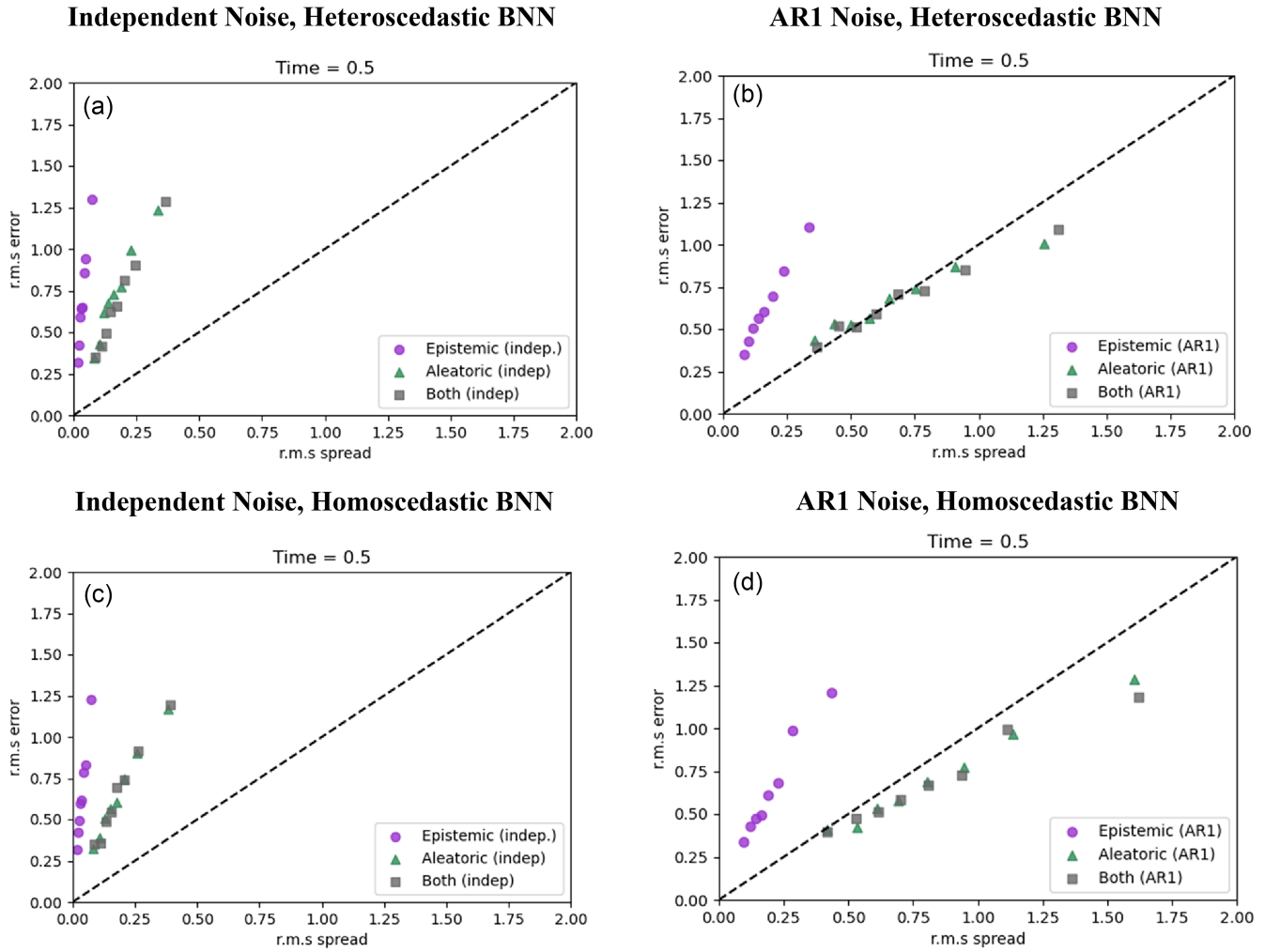

FIGURE 7 Root mean squared (r.m.s.) spread against r.m.s. error for (a) independent noise simulations and (b) AR1 simulations for the Bayesian neural network (BNN) that samples aleatoric (green), epistemic (purple) and both (grey) uncertainty. Each point represents a bin of 100 samples, over which the r.m.s. spread and r.m.s. error are calculated. The black dashed line shows the $y = x$ line, which represents a well-calibrated ensemble.

more sophisticated diagnostics beyond long-term probability distributions.

5.2 | Climate change scenarios

On climate time-scales, GCM users are typically interested in climate *change* given a change in forcing or emissions scenario. Here, we simulate a climate change experiment by perturbing the forcing, F, in Equation (3). The baseline climate that is used to generate the training data is $F = 20$ for all cases. Figure 8b shows the distribution over X for a decrease in forcing, where $F = 16$ and Figure 8c shows the distribution over X for an increase in forcing where $F = 24$. We do not see significant improvements in the distributions when moving from a deterministic forecast to a stochastic one that includes uncertainties. This suggests that on these time-scales, deterministic parameterisations may be sufficient.

5.3 | Perturbed-parameter ensembles

It is worth considering how epistemic uncertainty is usually treated by the climate modelling community. For a conventional parameterisation, parametric uncertainty is typically estimated through 'perturbed-parameter ensembles' (PPEs), where uncertainties on the parameters are defined based on domain knowledge. For each ensemble member, parameter values are sampled once and are fixed for the duration of the simulation. PPEs are used to quantify and, where possible, reduce parametric uncertainty by further constraining parameter values (Karmalkar *et al.*, 2019; Murphy *et al.*, 2007; Sexton *et al.*, 2021). Here,

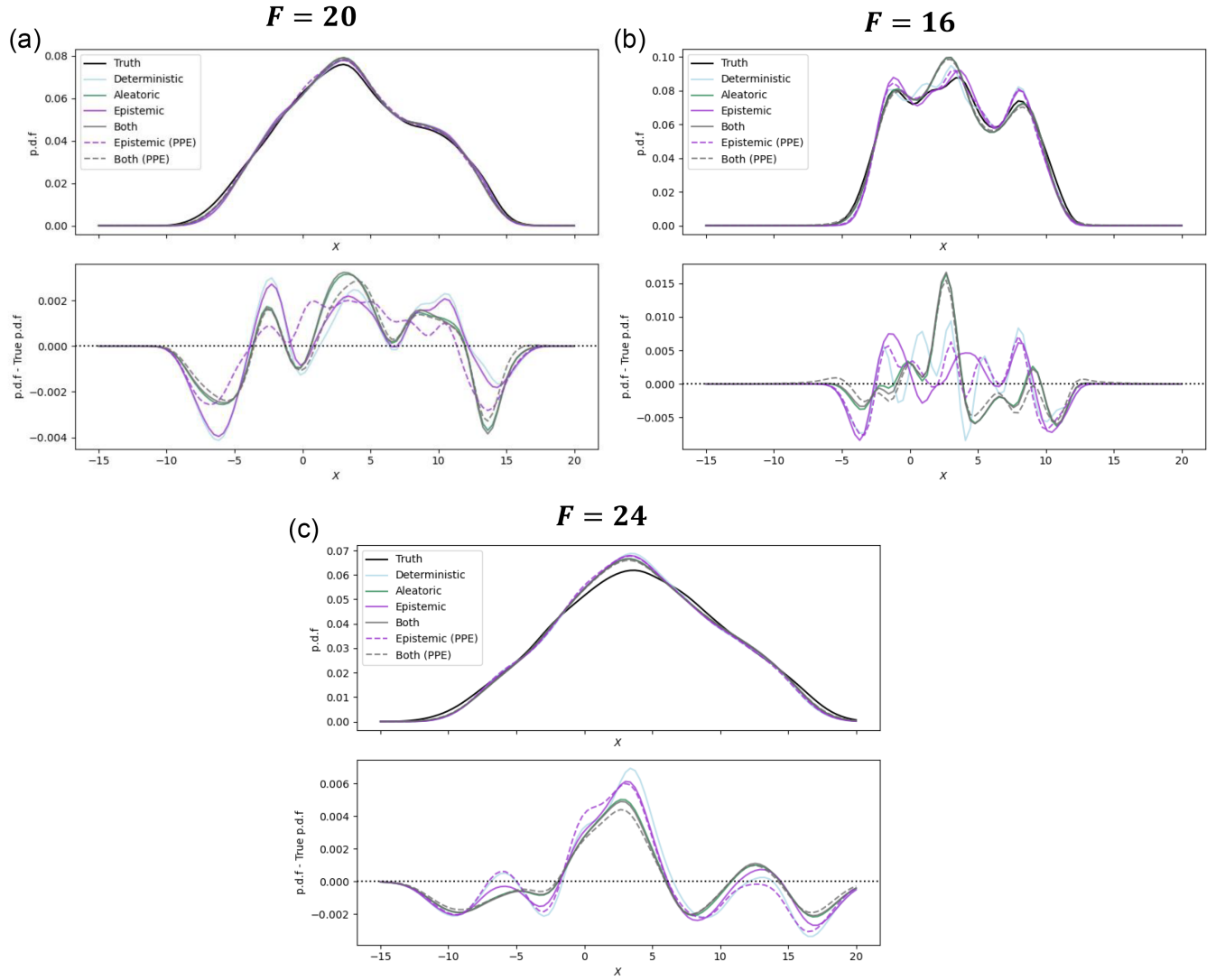

FIGURE 8 Probability distribution functions (p.d.f) and the difference between the p.d.f and the true p.d.f over X for longer integrations up to $T = 1000$ for (a) the 'baseline climate' with the same setup as used for the training data with $F = 20$, and perturbed climates with (b) $F = 16$, and (c) $F = 24$.

we carry out a PPE by sampling the BNN weights once for each ensemble member and holding them fixed throughout the simulation. We expect this to be a more realistic representation of epistemic uncertainty, which should remain constant across time-scales, rather than fluctuating on time-scales associated with correlations in the subgrid residual (as in the AR1 based approach). These simulations are also shown in Figure 8 by the purple dashed lines, but again, do not show consistent differences from the deterministic forecast. The grey dashed line shows simulations that combine a PPE with AR1 sampling for aleatoric uncertainty. These show slight improvements in capturing the tails of the distributions in the climate change simulations, but the differences are minor.

5.4 | Modes of variability

When considering long-term climate, we are usually concerned with obtaining the correct modes of variability. We typically measure these using scaler metrics, which either measure the phase of an oscillation or characterise phenomena using long-term statistics. Examples in the real world include the El Niño–Southern Oscillation (ENSO) occurring on time-scales of 3–5 years which is quantified by NINO3-4 indices (Wang *et al.*, 2017); the North Atlantic Oscillation (NAO) which is quantified by the NAO index (Hurrell *et al.*, 2003); and the Quasi-Biennial Oscillation (QBO) in stratospheric winds, which can exist in an easterly or westerly phase and can be characterised by their period or amplitude (Schenzinger *et al.*, 2017).

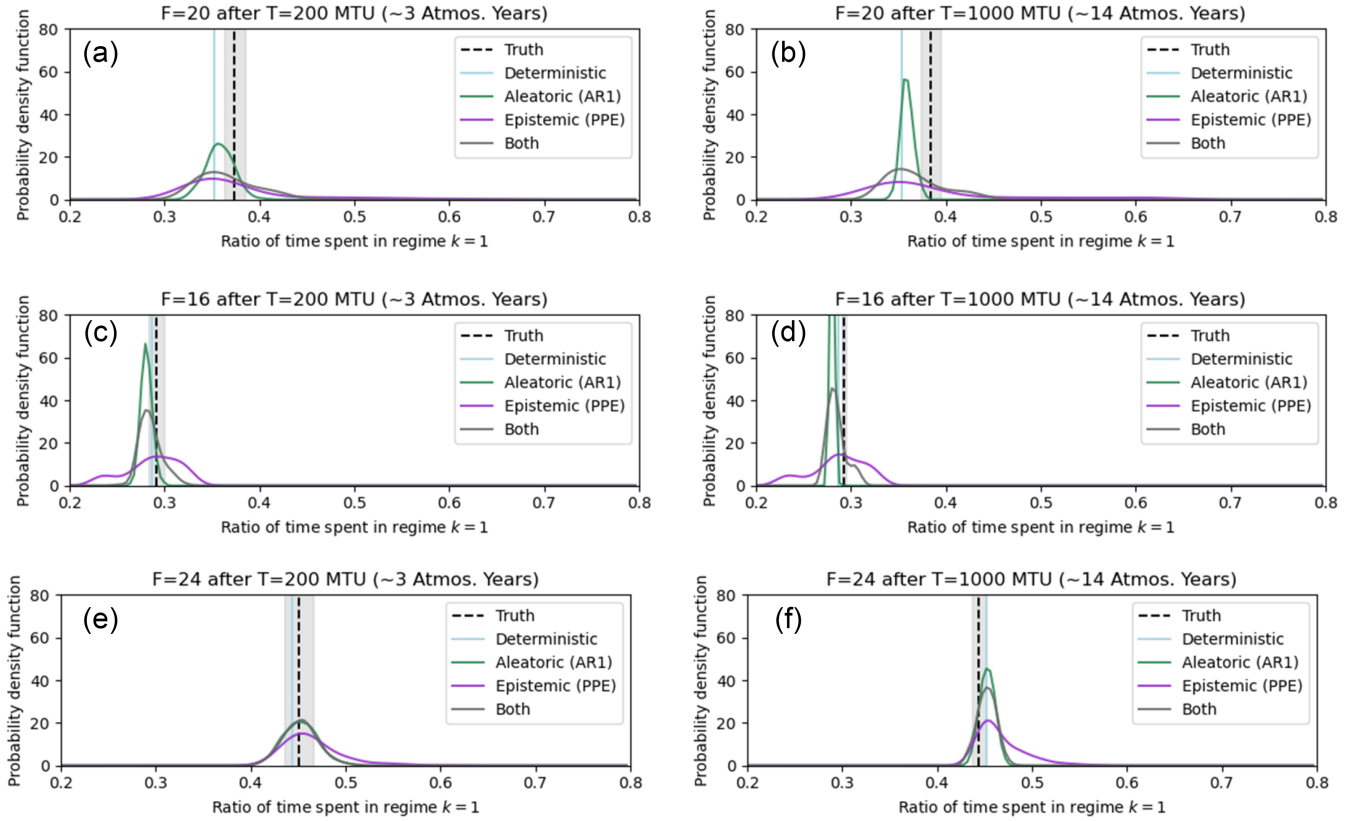

FIGURE 9 Distributions showing ratio of time spent in regime $k = 1$ across all ensemble members for (a,b) $F = 20$, (c,d) $F = 16$ and (e,f) $F = 24$. The black dashed line shows the truth and the grey shading around this shows the standard deviation across a 10-member ensemble of true simulations initialised with different initial conditions. The light blue line shows the mean from the deterministic prediction. The purple, green and grey lines show the distributions across the epistemic, aleatoric and both ensembles, calculated using kernel density estimation across all 50 ensemble members and 10 initial conditions. These are computed after (a,c,e) time $T = 200$ MTU (approximately three atmospheric years) and after (b,d,f) $T = 1000$ MTU (approximately 14 atmospheric years).

Lorenz noticed that the L96 model exhibits oscillatory modes with similar properties to these climate modes of variability (Lorenz, 2006). The spatial patterns over X over the circular domain can fall under two possible regimes: one which has a single wave (wavenumber $k = 1$) or one with two waves (wavenumber $k = 2$). Following the same method as Christensen *et al.* (2015b), we identify these regimes using principal component analysis (PCA, or empirical orthogonal functions, EOFs). Applying PCA decomposition over the truth time series highlights four modes of variability which explain 78% of the variance. The first two modes show a wavenumber $k = 1$ pattern which dominates 38% of the time, while the second two modes show a wavenumber $k = 2$ pattern which dominates 62% of the time. The two modes for each pattern exist because they are out of phase. We will compare whether the neural-network parameterisation simulations exhibit the same ratio of time spent in each regime.

We explore the PPE simulations here to see if parametric uncertainty can lead to systematic biases and different climate states. Here, we will consider a different climate state to be one that has different preference for either the $k = 1$ regime or the $k = 2$ regime. We run the simulations for T = 1000 MTU, each with 50 ensemble members. We also repeat this for 10 different initial conditions, to reduce the influence of the starting point.

Figure 9 shows the distributions across all ensemble members in the ratio of time spent in the $k = 1$ regime. The black dashed line shows the true ratio of time spent in the $k = 1$ regime, with the grey shading showing the standard deviation in this value across the 10 different initial conditions. When $F = 20$, Figure 9a shows a mismatch between the deterministic prediction and the truth, making any uncertainty estimate beneficial over a deterministic approach. The aleatoric uncertainty is narrower than epistemic uncertainty after time $T = 200$ (about three atmospheric years). As the simulation continues out to $T = 1000$ MTU (about 14 atmospheric years), Figure 9b shows that aleatoric uncertainty becomes even narrower, while the distribution representing the epistemic uncertainty remains similar. Here, the aleatoric simulations are overconfident, with the peak of

the distribution lying to the left of the truth, indicating most ensemble members underrepresent the time spent in the $k = 1$ regime.

This narrow aleatoric uncertainty arises because it reflects instantaneous subgrid variability which fluctuates rapidly, on time-scales associated with the AR1 process (∼0.33 MTU). These short-term fluctuations do not affect the long-term ratio of time spent in each regime. In contrast, fixed-parameter perturbations derived from epistemic uncertainty can induce persistent shifts in regime behaviour (Christensen *et al.*, 2015a). Some ensemble members remain in either the $k = 1$ or the $k = 2$ regime for extended periods. This leads to the large uncertainty that remains constant across the simulation. While this ensemble does capture the truth, the persistent regime behaviour is not realistic. The ensemble sampling both aleatoric and epistemic uncertainty shows that introducing stochastic variability from the aleatoric component can promote transitions between regimes, preventing members from remaining in a single state for too long. This shows that including short-term stochasticity can improve regime behaviour, consistent with results from GCM experiments (Dawson & Palmer, 2015). Together, these results highlight the importance of jointly considering both epistemic and aleatoric uncertainty.

Figure 9c–f shows the distributions for the climate change experiments, with Figure 9c,d corresponding to $F = 16$ and Figure 9e,f to $F = 24$. Although the BNN was trained on $F = 20$, it successfully reproduces the direction of the shift in regime preference under different forcings. This generalisation is encouraging for climate change applications, assuming the behaviour holds in a full GCM. As before, the aleatoric ensemble is overconfident and fails to capture the truth, particularly for $F = 16$. Here, introducing parameter perturbations increases ensemble diversity enough to capture the truth, indicating the potential benefit of including epistemic uncertainty in climate change experiments.

5.5 | Reducing model uncertainty

A key part of UQ lies in reducing model uncertainty in a process known as calibration. Calibration aims to constrain parameter values based on past observations and therefore focuses on reducing epistemic uncertainty. Here, we highlight a simple approach to calibration that uses the PPE we have generated. From the $F = 20$ simulations, at $T = 200$ MTU (about three atmospheric years), we select ensemble members that fall within one standard deviation of the truth (the grey shading in Figure 9a). This gives us six ensemble members from the epistemic ensemble, or 10 from the ensemble sampling both uncertainty types. These parameter choices are better constrained and less likely to produce persistent regime behaviour. Figure 10 shows the resulting constrained ensembles at $T = 1000$ MTU (∼14 atmospheric years) in bold for (Figure 10a,b) $F = 20$, (Figure 10c,d) $F = 16$, and (Figure 10e,f) $F = 24$. For (Figure 10a,b) $F = 20$, the constrained ensembles are centred on the truth and the spread is reduced significantly. This shows the potential to constrain parametric uncertainty once a parameterisation in coupled online, which could be particularly valuable for full GCMs. The constrained ensemble that samples both uncertainty types agrees particularly well with the truth (Figure 10b). This suggests an additional benefit of including aleatoric uncertainty during calibration, as it captures subgrid variability and reduces regime persistence. If reproduced in full GCMs, this approach could improve calibration of physics-based and ML parameterisations (Hourdin *et al.*, 2017; Williamson *et al.*, 2017).

For the climate change experiments, the epistemic distributions (Figure 10c for $F = 16$ and Figure 10e for $F = 24$) move closer to the truth with reduced uncertainty, even though they have been constrained on the $F = 20$ simulations. This demonstrates potential for improving climate model parameterisations through online fine-tuning of parameters. For the ensemble that samples both uncertainty types, when $F = 16$ (Figure 10d), the constrained ensemble slightly underestimates the truth, although it still provides an improvement upon the unconstrained ensembles, especially those that sample aleatoric uncertainty alone (Figure 9d). We do not see major improvements in the reduced ensemble that samples both for $F = 24$ (Figure 10f). Calibration based directly on the climate change simulations ($F = 16$, $F = 24$) is expected to improve performance and would be recommended for climate change applications in full GCMs, if available.

This simple approach highlights how we could potentially constrain parameter values based on observations and reduce model uncertainty. The constrained ensemble size could be reduced further, for instance, by constraining over longer periods of time or with stricter targets. This type of approach draws parallels with 'history matching', where PPEs are used to identify regions of the parameter space that agree best with observations in the past (Williamson *et al.*, 2013). History matching takes this further by using emulators that predict a target variable given the parameter values, allowing us to fully probe the parameter space and to predict another 'wave' of parameter values for the next PPE. The tuning is repeatedly carried out until reaching sufficient accuracy or until exhausting computational resources. History matching and other calibration techniques are usually designed for problems where there are $O(10)$ parameters. However, neural-network parameterisations typically have $O\left(10^5 - 10^7\right)$ parameters.

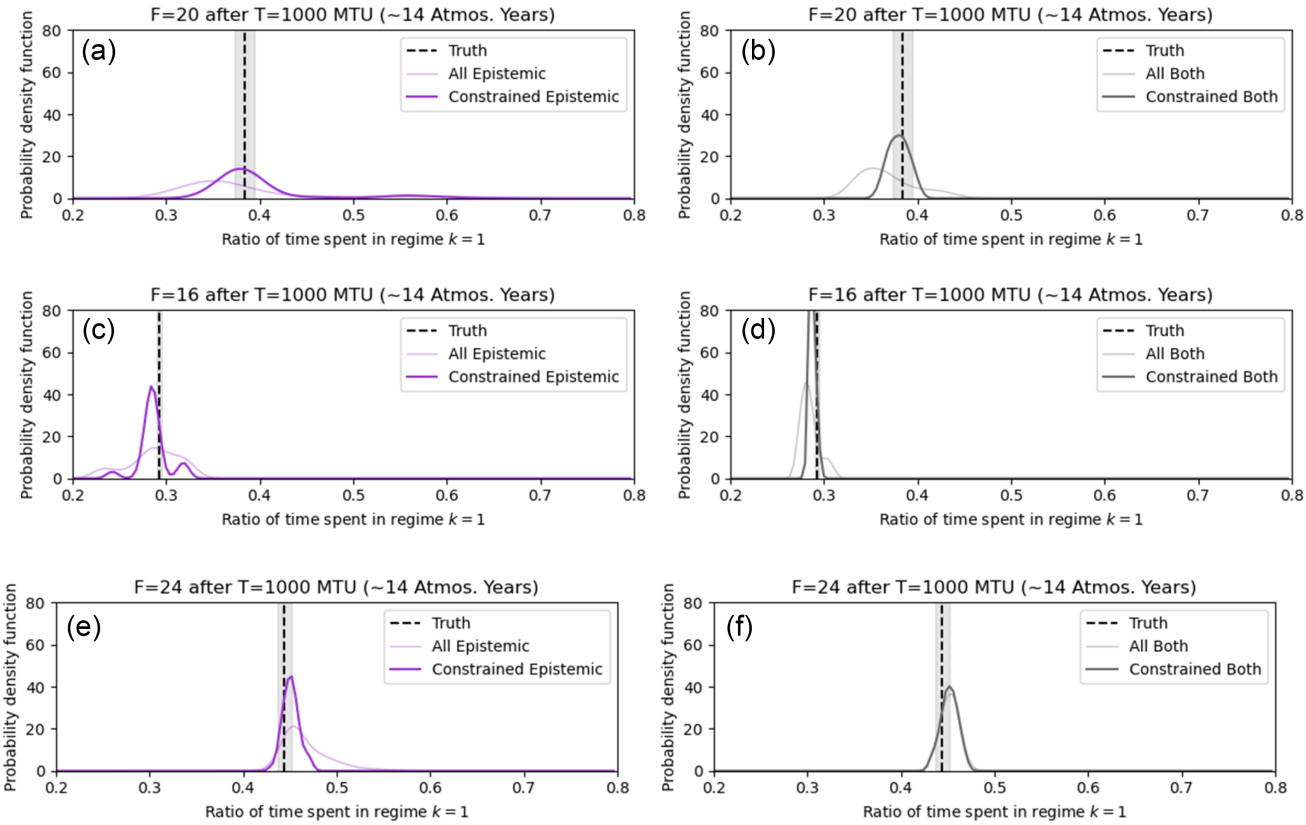

FIGURE 10 Distributions showing ratio of time spent in regime $k = 1$ across all ensemble members (light) and constrained ensemble members (bold) at $T = 1000$ MTU, for (a,b) $F = 20$, (c,d) $F = 16$ and (e,f) $F = 24$, where (a,c,e) show the epistemic ensembles and (b,d,f) the ensembles that sample both uncertainty types. The constrained ensembles are selected as ensemble members that agree with the truth to within one standard deviation after $T = 200$ MTU in the $F = 20$ simulations (grey shading in Figure 9a).

There is a need for further exploration of advanced calibration methods that constrain neural-network parameters once coupled online.

6 | DISCUSSION AND CONCLUSIONS

Using a Bayesian neural network, we have shown how we can identify epistemic (model) and aleatoric (data) uncertainties in parameterisations, including modelling aleatoric uncertainty as a function of the input data (heteroscedasticity). On short, weather time-scales, the dominant source of uncertainty is aleatoric uncertainty which arises from the subgrid variability. This validates stochastic machine-learning approaches that capture this form of uncertainty by training on probabilistic loss functions (e.g., Behrens *et al.*, 2025; Guillaumin & Zanna, 2021). On longer, climate time-scales, however, it is crucial to consider epistemic uncertainty. This is because including subgrid variability mostly influences short-term fluctuations and should not affect long-term climate statistics, whereas parameter choices remain fixed within a model and can alter the simulated climate. We find that model uncertainty remains approximately constant across long time-scales, in agreement with previous GCM studies (Hawkins & Sutton, 2009). Furthermore, neglecting parametric uncertainty can lead to overconfident responses to changing forcings and should therefore be accounted for in climate change simulations (e.g., Forster *et al.*, 2013; Murphy *et al.*, 2004).

We find that the approach to sampling uncertainty is also an important consideration when building stochastic parameterisations. For aleatoric uncertainty, including temporal correlations on time-scales associated with the subgrid variability, for instance through an AR1 process, is essential for producing well-calibrated ensembles in which spread and error are correlated (Arnold *et al.*, 2013). More reliable ensembles can also be achieved by representing aleatoric uncertainty as input-dependent (heteroscedasticity), which can be learned by the neural network. For epistemic uncertainty, keeping parameters fixed throughout the simulation better represents the constant parametric uncertainty we aim to capture (Hawkins & Sutton, 2009).

We also explored how PPEs can be used to fine-tune parameter values. While PPEs traditionally focus on $O(10)$

parameters, we extend this to a small neural network with about 500 parameters. We showed how the parameters can be crudely constrained based on observations and how this led to improved mean behaviour and reduced model uncertainty. This also held up under changing forcing experiments. Importantly, we found that including aleatoric uncertainty through the stochastic parameterisation improved the success of parameter calibration. If this behaviour holds in full GCMs, it could prove valuable for both physics-based and ML parameterisations, particularly given the potential for compensating errors that can arise during calibration – for example, from the interaction between resolved and unresolved atmospheric gravity waves (Cohen *et al.*, 2013; Mansfield & Sheshadri, 2022) or from competing radiative effects of clouds (Ma *et al.*, 2022; Zhao *et al.*, 2022). These challenges, often referred to as 'overtuning', can limit the broader adoption of calibration techniques (Hourdin *et al.*, 2017; Williamson *et al.*, 2017). To extend this research to full GCMs, more advanced calibration should be explored, such as history matching, ensemble Kalman methods, or Bayesian optimisation (Dunbar *et al.*, 2021; King *et al.*, 2024; Watson-Parris *et al.*, 2021; Williamson *et al.*, 2013). There is also the question of how to scale up PPE to neural-network parameterisations with $O\left(10^5 - 10^7\right)$. For this, it may be worth considering sampling approaches to improve computational efficiency and reduce redundancy in the ensemble, for instance by sampling the full parameter space or increasing parameter diversity (Karmalkar *et al.*, 2019; Sexton *et al.*, 2019, 2021).

This work provides insights into which types of uncertainty to target and how best to sample them across time-scales. Here, we used a simplified dynamical system as a toy model of the Earth's atmosphere. The next step is to explore how well this holds up in a more realistic GCM. In doing so, we expect that computational cost could become a bottleneck, since BNNs require frequent sampling of high-dimensional distributions and running large member ensembles to capture online uncertainties. Further exploration of cheaper approaches to capturing epistemic and aleatoric uncertainties may be required, such as evidential deep learning (Schreck *et al.*, 2024), informative priors for BNNs (Krishnan *et al.*, 2020), or Monte Carlo dropout (Gal & Ghahramani, 2016). The approaches outlined here serve as a starting point and we hope that this contributes to a growing emphasis on UQ for ML parameterisations.

Here, we have described a framework linking different uncertainties in weather and climate models to epistemic and aleatoric components (Table 1). However, it is possible that model parameters carry both forms of uncertainty. For instance, in convective parameterisations, parameters controlling entrainment can be influenced by unresolved variability (aleatoric) as well as lack of knowledge (epistemic) (e.g., Lock *et al.*, 2024). This blurs the distinction between these components. In weather and climate modelling, we typically use practical approaches to separate aleatoric and epistemic uncertainty, for instance, through stochastic perturbations to tendencies to capture subgrid variability (Buizza *et al.*, 1999) or through PPEs where parameter uncertainty is determined by domain knowledge (e.g., Sexton *et al.*, 2021). Methods developed in engineering could offer systematic ways to treat mixed forms of uncertainty, such as probability boxes which model both forms through a range of cumulative probability distributions (Bi *et al.*, 2023; Duran-Vinuesa & Cuervo, 2021) and stochastic kriging methods (Gaussian processes, J. Hu *et al.*, 2021). These approaches also have been shown to improve efficiency of model calibration (Bi *et al.*, 2023; Faber, 2005; Kiureghian & Ditlevsen, 2009).

Although aleatoric uncertainty alone may suffice for weather prediction, our results suggest that models designed for seamless prediction across time-scales (such as the Met Office Unified Model, Brown *et al.*, 2012) must sample both aleatoric and epistemic uncertainty. These results are obtained in a highly idealised system, yet they highlight fundamental mechanisms likely relevant to more complex models. If similar behaviour holds in full GCMs, then both sources of uncertainty should be incorporated into parameterisation development intended for use across time-scales. More generally, earth system prediction (whether fully data-driven models, hybrid GCMs, or entirely physics-based) should represent both aleatoric and epistemic uncertainty to capture the full range of possible outcomes across weather and climate regimes.

ACKNOWLEDGEMENTS

We are grateful to the editor of the journal and to the two anonymous reviewers and David John Gagne II for their insightful feedback which strengthened the manuscript. We are grateful to Yee Whye Teh for the insightful discussions. This research received support through Schmidt Sciences, LLC. H.M.C. was supported by the EERIE project (Grant Agreement No 101081383) funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Climate Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them. University of Oxford's contribution to EERIE is funded by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (grant number 10049639). H.M.C. was also supported by

the Leverhulme Trust Research Project Grant 'Exposing the nature of model error in weather and climate models' and through a Leverhulme Trust Research Leadership Award.

DATA AVAILABILITY STATEMENT

All code used to produce these simulations and all plots are available at https://github.com/lm2612/L96_UQ (Mansfield, 2026).

ORCID

Laura A. Mansfield https://orcid.org/0000-0002-6285-6045
Hannah M. Christensen https://orcid.org/0000-0001-8244-0218

REFERENCES

Arnold, H.M., Moroz, I.M. & Palmer, T.N. (2013) Stochastic parametrizations and model uncertainty in the Lorenz '96 system. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371, 20110479. Available from: https://doi.org/10.1098/rsta.2011.0479

Balwada, D., Abernathey, R., Acharya, S., Adcroft, A., Brener, J., Balaji, V. et al. (2024) Learning machine learning with Lorenz-96. *Journal of Open Source Education Education*, 7(82), 241. Available from: https://doi.org/10.21105/jose.00241

Barber, D. & Bishop, C. (1997) Ensemble Learning for Multi-Layer Networks. Advances in Neural Information Processing Systems 10. https://proceedings.neurips.cc/paper_files/paper/1997/hash/e816c635cad85a60fabd6b97b03cbcc9-Abstract.html.

Behrens, G., Beucler, T., Gentine, P., Iglesias-Suarez, F., Pritchard, M. & Eyring, V. (2022) Non-linear dimensionality reduction with a variational encoder decoder to understand convective processes in climate models. *Journal of Advances in Modeling Earth Systems*, 14(8), e2022MS003130. Available from: https://doi.org/10.1029/2022MS003130

Behrens, G., Beucler, T., Iglesias-Suarez, F., Yu, S., Gentine, P., Pritchard, M. et al. (2025) Simulating atmospheric processes in earth system models and quantifying uncertainties with deep learning multi-member and stochastic parameterizations. *Journal of Advances in Modeling Earth Systems*, 17(4), e2024MS004272. Available from: https://doi.org/10.1029/2024MS004272

Berner, J., Achatz, U., Batté, L., Bengtsson, L., Cámara, A.D.L., Christensen, H.M. et al. (2017) Stochastic parameterization: toward a new view of weather and climate models. *Bulletin of the American Meteorological Society*, 98(3), 565–588. Available from: https://doi.org/10.1175/BAMS-D-15-00268.1

Bi, S., Beer, M., Cogan, S. & Mottershead, J. (2023) Stochastic model updating with uncertainty quantification: an overview and tutorial. *Mechanical Systems and Signal Processing*, 204, 110784. Available from: https://doi.org/10.1016/j.ymssp.2023.110784

Blei, D.M., Kucukelbir, A. & McAuliffe, J.D. (2017) Variational inference: a review for statisticians. *Journal of the American Statistical Association*, 112(518), 859–877. Available from: https://doi.org/10.1080/01621459.2017.1285773

Bracco, A., Brajard, J., Dijkstra, H.A., Hassanzadeh, P., Lessig, C. & Monteleoni, C. (2025) Machine learning for the physics of climate. *Nature Reviews Physics*, 7(1), 6–20. Available from: https://doi.org/10.1038/s42254-024-00776-3

Brown, A., Milton, S., Cullen, M., Golding, B., Mitchell, J. & Shelly, A. (2012) Unified modeling and prediction of weather and climate: a 25-year journey. *Bulletin of the American Meteorological Society*, 93(12), 1865–1877. Available from: https://doi.org/10.1175/BAMS-D-12-00018.1

Buizza, R., Milleer, M. & Palmer, T.N. (1999) Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, 125(560), 2887–2908. Available from: https://doi.org/10.1002/qj.49712556006

Calvin, K., Dasgupta, D., Krinner, G., Mukherji, A., Thorne, P.W., Trisos, C. et al. (2023) IPCC, 2023: Climate Change 2023: Synthesis Report. In: Core Writing Team, Lee, H. & Romero, J. (Eds.) *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, pp. 1–34.

Carslaw, K.S., Lee, L.A., Pringle, K.J., Mann, G.W., Spracklen, D.V., Stier, P. et al. (2013) New approaches to quantifying the magnitude and causes of uncertainty in global aerosol models. *AIP Conference Proceedings*, 1527(1), 616–646. Available from: https://doi.org/10.1063/1.4803353

Chantry, M., Hatfield, S., Dueben, P., Polichtchouk, I. & Palmer, T. (2021) Machine learning emulation of gravity wave drag in numerical weather forecasting. *Journal of Advances in Modeling Earth Systems*, 13(7), e2021MS002477. Available from: https://doi.org/10.1029/2021MS002477

Chattopadhyay, A., Hassanzadeh, P. & Subramanian, D. (2020) Data-driven predictions of a multiscale Lorenz 96 chaotic system using machine-learning methods: reservoir computing, artificial neural network, and long short-term memory network. *Nonlinear Processes in Geophysics*, 27(3), 373–389. Available from: https://doi.org/10.5194/npg-27-373-2020

Christensen, H.M., Kouhen, S., Miller, G. & Parthipan, R. (2024) Machine learning for stochastic parametrization. *Environmental Data Science*, 3, e38. Available from: https://doi.org/10.1017/eds.2024.45

Christensen, H.M., Moroz, I.M. & Palmer, T.N. (2015a) Simulating weather regimes: impact of stochastic and perturbed parameter schemes in a simple atmospheric model. *Climate Dynamics*, 44(7), 2195–2214. Available from: https://doi.org/10.1007/s00382-014-2239-9

Christensen, H.M., Moroz, I.M. & Palmer, T.N. (2015b) Stochastic and perturbed parameter representations of model uncertainty in convection parameterization. *Journal of the Atmospheric Sciences*, 72(6), 2525–2544. Available from: https://doi.org/10.1175/JAS-D-14-0250.1

Christensen, H.M. & Zanna, L. (2022) Parametrization in weather and climate models. In: *Oxford research encyclopedia of climate science*. Oxford, UK: Oxford University Press.

Cloke, H.L. & Pappenberger, F. (2009). Ensemble flood forecasting: A review. *Journal of Hydrology*, 375(3-4), 613–626. Available from: https://doi.org/10.1016/j.jhydrol.2009.06.005

Cohen, N.Y., Gerber, E.P. & Bühler, O. (2013) Compensation between resolved and unresolved wave driving in the stratosphere: implications for downward control. *Journal of the Atmospheric Sciences*, 70(12), 3780–3798. Available from: https://doi.org/10.1175/JAS-D-12-0346.1

Couvreux, F., Hourdin, F., Williamson, D., Roehrig, R., Volodina, V., Villefranque, N. et al. (2021) Process-based climate model development harnessing machine learning: I. A calibration tool for parameterization improvement. *Journal of Advances in Modeling Earth Systems*, 13(3), e2020MS002217. Available from: https://doi.org/10.1029/2020MS002217

Dawson, A. & Palmer, T.N. (2015) Simulating weather regimes: impact of model resolution and stochastic parameterization. *Climate Dynamics*, 44(7), 2177–2193. Available from: https://doi.org/10.1007/s00382-014-2238-x

Dunbar, O.R.A., Garbuno-Inigo, A., Schneider, T. & Stuart, A.M. (2021) Calibration and uncertainty quantification of convective parameters in an idealized GCM. *Journal of Advances in Modeling Earth Systems*, 13(9), e2020MS002454. Available from: https://doi.org/10.1029/2020MS002454

Duran-Vinuesa, L. & Cuervo, D. (2021) Uncertainty quantification and propagation with probability boxes. *Nuclear Engineering and Technology*, 53(8), 2523–2533. Available from: https://doi.org/10.1016/j.net.2021.02.010

Eidhammer, T., Gettelman, A., Thayer-Calder, K., Watson-Parris, D., Elsaesser, G., Morrison, H. et al. (2024) An extensible perturbed parameter ensemble for the community atmosphere model version 6. *Geoscientific Model Development*, 17(21), 7835–7853. Available from: https://doi.org/10.5194/gmd-17-7835-2024

Espinosa, Z.I., Sheshadri, A., Cain, G.R., Gerber, E.P. & DallaSanta, K.J. (2022) Machine learning gravity wave parameterization generalizes to capture the QBO and response to increased CO2. *Geophysical Research Letters*, 49(8), e2022GL098174. Available from: https://doi.org/10.1029/2022GL098174

Eyring, V., Bony, S., Meehl, G.A., Senior, C.A., Stevens, B., Stouffer, R.J. et al. (2016) Overview of the coupled model intercomparison project phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, 9(5), 1937–1958. Available from: https://doi.org/10.5194/gmd-9-1937-2016

Faber, M.H. (2005) On the treatment of uncertainties and probabilities in engineering decision analysis. *Journal of Offshore Mechanics and Arctic Engineering*, 127(3), 243–248. Available from: https://doi.org/10.1115/1.1951776

Forster, P.M., Andrews, T., Good, P., Gregory, J.M., Jackson, L.S. & Zelinka, M. (2013) Evaluating adjusted forcing and model spread for historical and future scenarios in the CMIP5 generation of climate models. *Journal of Geophysical Research: Atmospheres*, 118(3), 1139–1150. Available from: https://doi.org/10.1002/jgrd.50174

Gagne, D.J., II, Christensen, H.M., Subramanian, A.C. & Monahan, A.H. (2020) Machine learning for stochastic parameterization: generative adversarial networks in the Lorenz '96 model. *Journal of Advances in Modeling Earth Systems*, 12(3), e2019MS001896. Available from: https://doi.org/10.1029/2019MS001896

Gal, Y. & Ghahramani, Z. (2016) Dropout as a Bayesian approximation: representing model uncertainty in deep learning. *Proceedings of the 33rd International Conference on Machine Learning*, 48, 1050–1059.

Gigerenzer, G., Hertwig, R., van den Broek, E., Fasolo, B. & Katsikopoulos, K.V. (2005) 'A 30% chance of rain tomorrow': how does the public understand probabilistic weather forecasts? *Risk Analysis*, 25(3), 623–629. Available from: https://doi.org/10.1111/j.1539-6924.2005.00608.x

Giles, D., Briant, J., Morcrette, C.J. & Guillas, S. (2024) Embedding machine-learnt sub-grid variability improves climate model precipitation patterns. *Communications Earth & Environment*, 5(1), 712. Available from: https://doi.org/10.1038/s43247-024-01885-8

Gneiting, T. & Katzfuss, M. (2014) Probabilistic forecasting. *Annual Review of Statistics and its Application*, 1(1), 125–151. Available from: https://doi.org/10.1146/annurev-statistics-062713-085831

Goan, E. & Fookes, C. (2020) Bayesian neural networks: an introduction and survey. In: Mengersen, K. Pudlo, P. & Robert, C. (Eds.) *Case Studies in Applied Bayesian Data Science. Lecture Notes in Mathematics*, Vol. 2259. Cham: Springer. Available from: https://doi.org/10.1007/978-3-030-42553-1_3

Grundner, A., Beucler, T., Gentine, P., Iglesias-Suarez, F., Giorgetta, M.A. & Eyring, V. (2022) Deep learning based cloud cover parameterization for ICON. *Journal of Advances in Modeling Earth Systems*, 14(12), e2021MS002959. Available from: https://doi.org/10.1029/2021MS002959

Guillaumin, A.P. & Zanna, L. (2021) Stochastic-deep learning parameterization of ocean momentum forcing. *Journal of Advances in Modeling Earth Systems*, 13(9), e2021MS002534. Available from: https://doi.org/10.1029/2021MS002534

Hawkins, E. & Sutton, R. (2009) The potential to narrow uncertainty in regional climate predictions. *Bulletin of the American Meteorological Society*, 90(8), 1095–1108. Available from: https://doi.org/10.1175/2009BAMS2607.1

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

Henn, B., Jauregui, Y.R., Clark, S.K., Brenowitz, N.D., McGibbon, J., Watt-Meyer, O. et al. (2024) A machine learning parameterization of clouds in a coarse-resolution climate model for unbiased radiation. *Journal of Advances in Modeling Earth Systems*, 16(3), e2023MS003949. Available from: https://doi.org/10.1029/2023MS003949

Heuer, H., Schwabe, M., Gentine, P., Giorgetta, M.A. & Eyring, V. (2024) Interpretable multiscale machine learning-based parameterizations of convection for ICON. *Journal of Advances in Modeling Earth Systems*, 16(8), e2024MS004398. Available from: https://doi.org/10.1029/2024MS004398

Hinton, G.E. & van Camp, D. (1993) Keeping neural networks simple. In: Gielen, S. & Kappen, B. (Eds.) *Internet Corporation for Assigned Names and Numbers (ICANN)'93*. London: Springer, pp. 11–18.

Hourdin, F., Mauritsen, T., Gettelman, A., Golaz, J.-C., Balaji, V., Duan, Q. et al. (2017) The art and science of climate model tuning. *Bulletin of the American Meteorological Society*, 98(3), 589–602. Available from: https://doi.org/10.1175/BAMS-D-15-00135.1

Hu, J., Zhou, Q., McKeand, A., Xie, T. & Choi, S.-K. (2021) A model validation framework based on parameter calibration under aleatory and epistemic uncertainty. *Structural and*

Multidisciplinary Optimization, 63(2), 645–660. Available from: https://doi.org/10.1007/s00158-020-02715-z

Hu, Z., Subramaniam, A., Kuang, Z., Lin, J., Yu, S., Hannah, W.M. et al. (2025) Stable machine-learning parameterization of subgrid processes in a comprehensive atmospheric model learned from embedded convection-permitting simulations. *Journal of Advances in Modeling Earth Systems*, 17(7), e2024MS004618. Available from: https://doi.org/10.1029/2024MS004618

Hüllermeier, E. & Waegeman, W. (2021) Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning*, 110(3), 457–506. Available from: https://doi.org/10.1007/s10994-021-05946-3

Hurrell, J.W., Kushnir, Y., Ottersen, G. & Visbeck, M. (2003) An overview of the North Atlantic oscillation. In: *The North Atlantic oscillation: climatic significance and environmental impact*. Washington, DC: American Geophysical Union, pp. 1–35.

Karmalkar, A.V., Sexton, D.M.H., Murphy, J.M., Booth, B.B.B., Rostron, J.W. & McNeall, D.J. (2019) Finding plausible and diverse variants of a climate model. part II: development and validation of methodology. *Climate Dynamics*, 53(1), 847–877. Available from: https://doi.org/10.1007/s00382-019-04617-3

Kendall, A. & Gal, Y. (2017) What uncertainties do we need in Bayesian deep learning for computer vision? In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S. et al. (Eds.) *Advances in neural information processing systems*, Vol. 30. Long Beach, CA: Curran Associates Inc.

King, R.C., Mansfield, L.A. & Sheshadri, A. (2024) Bayesian history matching applied to the calibration of a gravity wave parameterization. *Journal of Advances in Modeling Earth Systems*, 16(4), e2023MS004163. Available from: https://doi.org/10.1029/2023MS004163

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

Krishnan, R., Subedar, M. & Tickoo, O. (2020) Specifying weight priors in bayesian deep neural networks with empirical bayes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(4), 4477–4484.

Lang, S.T.K., Lock, S.-J., Leutbecher, M., Bechtold, P. & Forbes, R.M. (2021) Revision of the stochastically perturbed Parametrisations model uncertainty scheme in the integrated forecasting system. *Quarterly Journal of the Royal Meteorological Society*, 147(735), 1364–1381. Available from: https://doi.org/10.1002/qj.3978

Leutbecher, M. (2009) Diagnosis of Ensemble Forecasting Systems.

Leutbecher, M. & Palmer, T.N. (2008) Ensemble forecasting. *Journal of Computational Physics*, 227(7), 3515–3539. Available from: https://doi.org/10.1016/j.jcp.2007.02.014

Lock, A.P., Whitall, M., Stirling, A.J., Williams, K.D., Lavender, S.L., Morcrette, C. et al. (2024) The performance of the CoMorph-A convection package in global simulations with the met Office unified model. *Quarterly Journal of the Royal Meteorological Society*, 150(763), 3527–3543. Available from: https://doi.org/10.1002/qj.4781

Lorenz, E.N. (2006) Predictability – a problem partly solved. In: Hagedorn, R. & Palmer, T. (Eds.) *Predictability of weather and climate*. Cambridge, UK: Cambridge University Press, pp. 40–58.

Ma, P.-L., Harrop, B.E., Larson, V.E., Neale, R.B., Gettelman, A., Morrison, H. et al. (2022) Better calibration of cloud parameterizations and subgrid effects increases the fidelity of the E3SM atmosphere model version 1. *Geoscientific Model Development*, 15(7), 2881–2916. Available from: https://doi.org/10.5194/gmd-15-2881-2022

Mansfield, L.A. (2026) L96_UQ [Computer software].

Mansfield, L.A. & Sheshadri, A. (2022) Calibration and uncertainty quantification of a gravity wave parameterization: a case study of the quasi-biennial oscillation in an intermediate complexity climate model. *Journal of Advances in Modeling Earth Systems*, 14(11), e2022MS003245. Available from: https://doi.org/10.1029/2022MS003245

McGovern, A., Bostrom, A., Davis, P., Demuth, J.L., Ebert-Uphoff, I., He, R. et al. (2022) NSF AI Institute for Research on trustworthy AI in weather, climate, and coastal oceanography (AI2ES). *Bulletin of the American Meteorological Society*, 103(7), E1658–E1668. Available from: https://doi.org/10.1175/BAMS-D-21-0020.1

Miller, G.A., Stier, P. & Christensen, H.M. (2025) Characterizing uncertainty in deep convection triggering using explainable machine learning. *Journal of the Atmospheric Sciences*, 82(6), 1093–1111. Available from: https://doi.org/10.1175/JAS-D-24-0085.1

Morcrette, C., Cave, T., Reid, H., da Silva Rodrigues, J., Deveney, T., Kreusser, L. et al. (2025) Scale-aware parameterization of cloud fraction and condensate for a global atmospheric model machine-learned from coarse-grained kilometer-scale simulations. *Journal of Advances in Modeling Earth Systems*, 17(4), e2024MS004651. Available from: https://doi.org/10.1029/2024MS004651

Murphy, J.M., Booth, B.B.B., Collins, M., Harris, G.R., Sexton, D.M.H. & Webb, M.J. (2007) A methodology for probabilistic predictions of regional climate change from perturbed physics ensembles. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 365(1857), 1993–2028. Available from: https://doi.org/10.1098/rsta.2007.2077

Murphy, J.M., Sexton, D.M.H., Barnett, D.N., Jones, G.S., Webb, M.J., Collins, M. et al. (2004) Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature*, 430(7001), 768–772. Available from: https://doi.org/10.1038/nature02771

Nadiga, B.T., Sun, X. & Nash, C. (2022) Stochastic parameterization of column physics using generative adversarial networks. *Environmental Data Science*, 1, e22. Available from: https://doi.org/10.1017/eds.2022.32

Palmer, T.N. (2019) Stochastic weather and climate models. *Nature Reviews Physics*, 1(7), 7. Available from: https://doi.org/10.1038/s42254-019-0062-2

Parthipan, R., Christensen, H.M., Hosking, J.S. & Wischik, D.J. (2023) Using probabilistic machine learning to better model temporal patterns in parameterizations: a case study with the Lorenz 96 model. *Geoscientific Model Development*, 16(15), 4501–4519. Available from: https://doi.org/10.5194/gmd-16-4501-2023

Perezhogin, P., Zanna, L. & Fernandez-Granda, C. (2023) Generative data-driven approaches for stochastic subgrid parameterizations in an idealized ocean model (arXiv:2302.07984). arXiv. http://arxiv.org/abs/2302.07984

Ranganath, R., Gerrish, S. & Blei, D.M. (2013) Black Box Variational Inference (arXiv:1401.0118). arXiv.

Raoult, N., Beylat, S., Salter, J.M., Hourdin, F., Bastrikov, V., Ottlé, C. et al. (2024) Exploring the potential of history matching for

land surface model calibration. *Geoscientific Model Development*, 17(15), 5779–5801. Available from: https://doi.org/10.5194/gmd-17-5779-2024

Rasp, S. (2020) Coupled online learning as a way to tackle instabilities and biases in neural network parameterizations: general algorithms and Lorenz 96 case study (v1.0). *Geoscientific Model Development*, 13(5), 2185–2196. Available from: https://doi.org/10.5194/gmd-13-2185-2020

Rasp, S., Pritchard, M.S. & Gentine, P. (2018) Deep learning to represent subgrid processes in climate models. *Proceedings of the National Academy of Sciences of the United States of America*, 115, 9684–9689. Available from: https://doi.org/10.1073/pnas.1810286115

Ross, A., Li, Z., Perezhogin, P., Fernandez-Granda, C. & Zanna, L. (2023) Benchmarking of machine Learning Ocean subgrid parameterizations in an idealized model. *Journal of Advances in Modeling Earth Systems*, 15(1), e2022MS003258. Available from: https://doi.org/10.1029/2022MS003258

Rougier, J. (2007) Probabilistic inference for future climate using an Ensemble of Climate Model Evaluations. *Climatic Change*, 81(3), 247–264. Available from: https://doi.org/10.1007/s10584-006-9156-9

Roberts, G.O., Gelman, A. & Gilks, W.R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability*, 7(1), 110–120. Available from:. https://doi.org/10.1214/aoap/1034625254

Schenzinger, V., Osprey, S., Gray, L. & Butchart, N. (2017) Defining metrics of the quasi-biennial oscillation in global climate models. *Geoscientific Model Development*, 10(6), 2157–2168. Available from: https://doi.org/10.5194/gmd-10-2157-2017

Schreck, J.S., Gagne, D.J., Becker, C., Chapman, W.E., Elmore, K., Fan, D. et al. (2024) Evidential deep learning: enhancing predictive uncertainty estimation for earth system science applications. *Artificial Intelligence for the Earth Systems*, 3(4), 230093. Available from: https://doi.org/10.1175/AIES-D-23-0093.1

Sengupta, K., Pringle, K., Johnson, J.S., Reddington, C., Browse, J., Scott, C.E. et al. (2021) A global model perturbed parameter ensemble study of secondary organic aerosol formation. *Atmospheric Chemistry and Physics*, 21(4), 2693–2723. Available from: https://doi.org/10.5194/acp-21-2693-2021

Sexton, D.M.H., Karmalkar, A.V., Murphy, J.M., Williams, K.D., Boutle, I.A., Morcrette, C.J. et al. (2019) Finding plausible and diverse variants of a climate model. part 1: establishing the relationship between errors at weather and climate time scales. *Climate Dynamics*, 53(1), 989–1022. Available from: https://doi.org/10.1007/s00382-019-04625-3

Sexton, D.M.H., McSweeney, C.F., Rostron, J.W., Yamazaki, K., Booth, B.B.B., Murphy, J.M. et al. (2021) A perturbed parameter ensemble of HadGEM3-GC3.05 coupled model projections: part 1: selecting the parameter combinations. *Climate Dynamics*, 56(11), 3395–3436. Available from: https://doi.org/10.1007/s00382-021-05709-9

Shutts, G. (2005) A kinetic energy backscatter algorithm for use in ensemble prediction systems. *Quarterly Journal of the Royal Meteorological Society*, 131(612), 3079–3102. Available from: https://doi.org/10.1256/qj.04.106

Slingo, J. & Palmer, T. (2011) Uncertainty in weather and climate prediction. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, 369, 4751–4767. Available from: https://doi.org/10.1098/rsta.2011.0161

Souza, A.N., Wagner, G.L., Ramadhan, A., Allen, B., Churavy, V., Schloss, J. et al. (2020) Uncertainty quantification of ocean parameterizations: application to the K-profile-parameterization for penetrative convection. *Journal of Advances in Modeling Earth Systems*, 12(12), e2020MS002108. Available from: https://doi.org/10.1029/2020MS002108

Stainforth, D.A., Aina, T., Christensen, C., Collins, M., Faull, N., Frame, D.J. et al. (2005) Uncertainty in predictions of the climate response to rising levels of greenhouse gases. *Nature*, 433(7024), 7024. Available from: https://doi.org/10.1038/nature03301

Ukkonen, P. (2022) Exploring pathways to more accurate machine learning emulation of atmospheric radiative transfer. *Journal of Advances in Modeling Earth Systems*, 14(4), e2021MS002875. Available from: https://doi.org/10.1029/2021MS002875

Ukkonen, P. & Chantry, M. (2025) Vertically recurrent neural networks for sub-grid parameterization. *Journal of Advances in Modeling Earth Systems*, 17(6), e2024MS004833. Available from: https://doi.org/10.1029/2024MS004833

Valdenegro-Toro, M. & Mori, D.S. (2022) A deeper look into aleatoric and epistemic uncertainty disentanglement. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. New Orleans, LA: IEEE, pp. 1508–1516.

Wang, C., Deser, C., Yu, J.-Y., DiNezio, P. & Clement, A. (2017) El Niño and southern oscillation (ENSO): a review. In: Glynn, P.W., Manzello, D.P. & Enochs, I.C. (Eds.) *Coral reefs of the eastern tropical Pacific: persistence and loss in a dynamic environment*. Dordrecht: Springer Netherlands, pp. 85–106.

Watson-Parris, D., Williams, A., Deaconu, L. & Stier, P. (2021) Model calibration using ESEm v1.1.0—an open, scalable earth system emulator. *Geoscientific Model Development*, 14(12), 7659–7672. Available from: https://doi.org/10.5194/gmd-14-7659-2021

Watt-Meyer, O., Brenowitz, N.D., Clark, S.K., Henn, B., Kwa, A., McGibbon, J. et al. (2024) Neural network parameterization of subgrid-scale physics from a realistic geography global storm-resolving simulation. *Journal of Advances in Modeling Earth Systems*, 16(2), e2023MS003668. Available from: https://doi.org/10.1029/2023MS003668

Wilks, D.S. (2005) Effects of stochastic parametrizations in the Lorenz '96 system. *Quarterly Journal of the Royal Meteorological Society*, 131(606), 389–407. Available from: https://doi.org/10.1256/qj.04.03

Wilks, D.S. (2006) *Statistical methods in the atmospheric sciences*, 2nd edition. London: Academic Press.

Williamson, D.B., Blaker, A.T. & Sinha, B. (2017) Tuning without over-tuning: parametric uncertainty quantification for the NEMO ocean model. *Geoscientific Model Development*, 10(4), 1789–1816. Available from: https://doi.org/10.5194/gmd-10-1789-2017

Williamson, D.B., Goldstein, M., Allison, L., Blaker, A., Challenor, P., Jackson, L. et al. (2013) History matching for exploring and reducing climate model parameter space using observations and a large perturbed physics ensemble. *Climate Dynamics*, 41(7–8), 1703–1729. Available from: https://doi.org/10.1007/s00382-013-1896-4

Yu, S., Hu, Z., Subramaniam, A., Hannah, W., Peng, L., Lin, J. et al. (2024) *ClimSim-online: a large multi-scale dataset and framework for hybrid ML-physics climate emulation* (arXiv:2306.08754). arXiv. http://arxiv.org/abs/2306.08754

Yuval, J. & O'Gorman, P.A. (2023) Neural-network parameterization of subgrid momentum transport in the atmosphere. *Journal of Advances in Modeling Earth Systems*, 15(4), e2023MS003606. Available from: https://doi.org/10.1029/2023MS003606

Zhang, C., Perezhogin, P., Gultekin, C., Adcroft, A., Fernandez-Granda, C. & Zanna, L. (2023) *Implementation and evaluation of a machine learned mesoscale Eddy parameterization into a Numerical Ocean circulation model* (arXiv:2303.00962). arXiv. http://arxiv.org/abs/2303.00962

Zhao, L., Wang, Y., Zhao, C., Dong, X. & Yung, Y.L. (2022) Compensating errors in cloud radiative and physical properties over the Southern Ocean in the CMIP6 climate models. *Advances in Atmospheric Sciences*, 39(12), 2156–2171. Available from: https://doi.org/10.1007/s00376-022-2036-z

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Mansfield, L.A. & Christensen, H.M. (2026) Epistemic and aleatoric uncertainty quantification in weather and climate models. *Quarterly Journal of the Royal Meteorological Society*, e70219. Available from: https://doi.org/10.1002/qj.70219