



Research



Cite this article: Schmid N, Fernandes del Pozo D, Waegeman W, Hasenauer J. 2025 Assessment of uncertainty quantification in universal differential equations. *Phil. Trans. R. Soc. A* **383**: 20240444.

<https://doi.org/10.1098/rsta.2024.0444>

Received: 2 October 2024

Accepted: 13 December 2024

One contribution of 13 to a theme issue
'Uncertainty quantification for healthcare and
biological systems (Part 2)'.

Subject Areas:

computational biology

Keywords:

uncertainty quantification, universal
differential equations, scientific machine
learning

Author for correspondence:

Jan Hasenauer

e-mail: jan.hasenauer@uni-bonn.de

† These authors contributed equally to the
study.

Electronic supplementary material is available
online at: [https://doi.org/10.6084/m9.
figshare.c.7718169](https://doi.org/10.6084/m9.figshare.c.7718169)

Assessment of uncertainty
quantification in universal
differential equations

Nina Schmid¹, David Fernandes del Pozo²,
Willem Waegeman^{2,†} and Jan Hasenauer^{1,3,†}

¹Life & Medical Sciences (LIMES) Institute, University of Bonn, Bonn, Germany

²Department of Data Analysis and Mathematical Modelling, Ghent University,
Ghent, Belgium

³Helmholtz Center Munich, German Research Center for Environmental Health,
Computational Health Center, Munich, Germany

JH, 0000-0002-4935-3312

Scientific machine learning is a new class of approaches that integrate physical knowledge and mechanistic models with data-driven techniques to uncover the governing equations of complex processes. Among the available approaches, universal differential equations (UDEs) combine prior knowledge in the form of mechanistic formulations with universal function approximators, such as neural networks. Integral to the efficacy of UDEs is the joint estimation of parameters for both the mechanistic formulations and the universal function approximators using empirical data. However, the robustness and applicability of these resultant models hinge upon the rigorous quantification of uncertainties associated with their parameters and predictive capabilities. In this work, we provide a formalization of uncertainty quantification (UQ) for UDEs and investigate key frequentist and Bayesian methods. By analyzing three synthetic examples of varying complexity, we evaluate the validity and efficiency of ensembles, variational inference and Markov-chain Monte Carlo sampling as epistemic UQ methods for UDEs.

This article is part of the theme issue 'Uncertainty quantification for healthcare and biological systems (Part 2)'.

1. Introduction

Two primary paradigms govern the modelling of dynamical systems: (i) Mechanistic modelling, which builds on first principles to derive model structures and informs parameters using data; and (ii) machine learning (ML)-based modelling, which constructs governing equations directly from data. Prime examples of ML-based modelling of dynamical systems include composite artificial neural networks (ANNs) based on simple integrator schemes [1] and Runge–Kutta neural networks (RKNNs) [2]. This groundbreaking work led to the development of general neural ordinary differential equations (NODEs) [3]. Yet, while NODEs allow for uncovering governing equations of dynamical systems, they do not easily facilitate the integration of prior knowledge.

Scientific machine learning (SciML) unites the paradigms of mechanistic and ML-based modelling [4]. For instance, physics-informed neural networks (PINNs) describe solutions of dynamical systems using customized loss functions that favour predictions in agreement with first-principles mechanistic models [5]. Universal differential equations (UDEs) model the time derivative of the state of a system using a combination of terms derived from first principles and flexible ANNs [6]. In our opinion, UDEs stand out because they allow for the integration of prior knowledge and hard physical constraints (e.g. mass conservation or boundedness) [7], thus facilitating generalization. Moreover, they provide governing equations that can be further interrogated.

The parameters of both the mechanistic and the neural network components of a UDE are jointly estimated using data. Interpretation of modelling results hinges on quantifying uncertainties, encompassing mechanistic parameter values and predictions for the entire model or its components. Uncertainty quantification (UQ) of parameters is crucial because it provides insights into the reliability and range of potential values, allowing researchers to understand the robustness and credibility of their model's mechanistic foundations. UQ of the prediction is equally important, as it offers a measure of the model's reliability, e.g. for perturbation studies or scenario analysis, aiding decision-making by acknowledging the inherent uncertainty in forecasting outcomes.

UQ is a highly researched topic for both dynamical mechanistic modelling [8] and machine learning [9,10]. First-order uncertainty describes the inherent and irreducible stochasticity of the predictions (aleatoric uncertainty), while second-order uncertainty describes uncertainty originating from the uncertainty of parameter estimates (which is one component of epistemic uncertainty—see §2 (c)). By choosing a suitable noise model, the assessment of aleatoric uncertainty is well-defined. Hence, in this work, while reporting results on aleatoric uncertainty, we focus on estimating epistemic uncertainty.

In supervised machine learning, various methods aim to quantify epistemic uncertainty, and many of these methods have a resemblance to the field of mechanistic modelling. A fully Bayesian perspective is realized by Markov-Chain Monte Carlo (MCMC) sampling methods [11] and approximation methods like Variational Inference [12], yielding parameter distributions instead of point estimates. Deep ensembles [13] and multi-start ensembles in dynamic modelling [8,14] are both randomization-based ensemble approaches. Key differences between mechanistic modelling and machine learning are the number and interpretability of the parameters. Hence, some flavours of UQ methods are exclusively used in deep learning, like dropout as a Bayesian approximation [15]. Others are more common in dynamic modelling, like Profile Likelihood (PL) calculation [16] or asymptotic confidence intervals via the Fisher Information Matrix (FIM) [17].

For some modelling approaches in the field of SciML, like PINNs [18], a thorough investigation of UQ exists [19]. In contrast to other methods, UDEs embed neural networks directly in the differential equations. While this allows the incorporation of arbitrary levels of prior knowledge, it also yields challenges like over-parametrized differential equations with correlated parameters in combination with numerically more challenging simulations. To the best of our knowledge, previous work explored only basic UQ implementations for multi-start-optimization [20] or Bayesian neural networks [21,22] and only considered fully observed or densely measured state variables.

In this paper, we present several key contributions. Firstly, we introduce a formal definition of uncertainty tailored to UDEs, aiming to enhance precision and applicability in uncertainty

assessments within this framework (§2). Secondly, we conduct an in-depth discussion of current epistemic UQ methods applicable to UDEs (§3). Lastly, we evaluate and compare the performance of a diverse set of UQ methods by investigating three synthetic examples (§4). Each synthetic example is implemented using several noise models, yielding eleven data scenarios in total. Synthetically generated data allows us to compare the methods' results with an underlying ground truth. Our investigation spans considerations of computing time, estimations of aleatoric and epistemic uncertainty, parameter and prediction uncertainty, and different noise models, encompassing continuous and discrete distributions. Although the groups of UQ methods have been investigated before [8,23], we find novel insights in the context of UDEs.

2. Formalizing precision: a tailored definition of uncertainty for UDEs

In the following subsections, we first define the general setup of dynamic models, formally introduce UDEs and conclude by presenting different sources of uncertainty and discussing their relevance for UDEs.

(a) Dynamic models

Let $\mathbf{x}(t) \in \mathbb{R}^{n_x}$ be a time-dependent variable, denoting the state of a system at time t , that can be represented using a dynamic model. Dynamic models describe the value of \mathbf{x} by parameterizing the derivative of $\mathbf{x}(t)$ and its initial condition $\mathbf{x}(t_0)$ using a vector field $f : \mathbb{R} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_\theta} \rightarrow \mathbb{R}^{n_x}$:

$$\frac{d\mathbf{x}}{dt} = f(t, \mathbf{x}, \boldsymbol{\theta}_f), \quad \mathbf{x}(t_0) = \mathbf{x}_0, \quad (2.1)$$

where $\boldsymbol{\theta}_f \in \mathbb{R}^{n_\theta}$ and $\mathbf{x}_0 \in \mathbb{R}^{n_x}$ are model parameters and initial conditions, respectively. Often, f is unknown and an estimate \hat{f} is used instead. Let $\hat{\boldsymbol{\theta}}_f$ and $\hat{\mathbf{x}}_0$ be the parameters and initial conditions of \hat{f} which we estimate based on n_t discrete measurements at time points $\{t_1, t_2, \dots, t_{n_t}\}$.

In many real-life scenarios, the state variables cannot be measured directly. Accordingly, the prediction of the differential equation model needs to be transformed using an observable function h to values predicting the measurable observables as:

$$\hat{\mathbf{y}}(t) = h(\hat{\mathbf{x}}(t)) \in \mathbb{R}^{n_y} \quad \text{with} \quad \hat{\mathbf{x}}(t) = \int_{t_0}^t \hat{f}(s, \hat{\mathbf{x}}(s), \hat{\boldsymbol{\theta}}_f) ds + \hat{\mathbf{x}}_0. \quad (2.2)$$

Here, $\hat{\mathbf{x}}(t)$ is the estimate of $\mathbf{x}(t)$ that we get from using \hat{f} , $\hat{\boldsymbol{\theta}}_f$ and $\hat{\mathbf{x}}_0$. An example for h comes from infectious disease modelling, where we often observe infections, but, e.g. not the number of susceptible, exposed or recovered persons.

Furthermore, measurements are subject to noise. Instead of measuring the underlying true value $\bar{\mathbf{y}}(t_k) = h(\mathbf{x}(t_k))$ of the observable, we observe the random variable $\mathbf{y}(t_k) \sim P$ with P being a probability distribution. In general, we do not know the underlying distribution of $\mathbf{y}(t_k)$. Instead, we fit a parametric distribution, which is called the noise model. There exist different formulations for noise models, with the Gaussian being the most prominent representative. Depending on the characteristics of the underlying measurements like discreteness, overdispersion or skewness, other noise formulations may be more suitable. In the present work, we will focus on two commonly used noise models in the context of infectious disease modelling [24,25], the Gaussian noise model for continuous data and the Negative Binomial noise model for overdispersed and discrete data:

- **Gaussian noise model:** Let $\epsilon(t_k) \sim \mathcal{N}(0, \sigma^2 I)$. Then, we observe $\mathbf{y}(t_k) = \bar{\mathbf{y}}(t_k) + \epsilon(t_k)$, where σ is the constant standard deviation of the Gaussian distribution.
- **Negative Binomial noise model:** Let $\mathbf{y} \in \mathbb{R}^{n_y}$. The observed variable y_i follows a Negative Binomial distribution with mean $\bar{y}_i(t_k)$ and dispersion parameter d , i.e. $y_i(t_k) \sim \text{NegBin}(\bar{y}_i(t_k), d)$, for all $i \in \{1, \dots, n_y\}$.

In both cases, we assume that the i.i.d. assumption holds. Let $\theta = \{\theta_f, \theta_{np}\}$, where θ_{np} is the noise parameter of the respective noise model and $p(\mathbf{y}(t)|\theta)$ the probability density function with mean value $\hat{\mathbf{y}}(t)$. Then, the objective of the optimization process is to maximize the likelihood of observing the data $\mathcal{D} = \{(t_i, \mathbf{y}(t_i)) | i = 1, \dots, n_t\}$ given the parameters θ .

(b) Universal differential equations

UDEs combine known mechanistic terms f_{mech} with universal function approximators (in this work neural networks) f_{net} to describe the right-hand side of equation (2.1) [6]. For instance, the neural network can be used to describe the time-varying input of an otherwise purely mechanistic ordinary differential equation, i.e. for a fixed t we have $\hat{f}(t, \mathbf{x}, \theta) = \hat{f}_{\text{mech}}(t, \mathbf{x}, \theta_f)$, with $\theta_f = (\theta_{\text{mech}}, \hat{f}_{\text{net}}(t, \theta_{\text{net}}))$. Alternatively, it can describe individual terms of the state derivatives, e.g. $\hat{f}(t, \mathbf{x}, \theta) = \hat{f}_{\text{mech}}(t, \mathbf{x}, \theta_{\text{mech}}) + \hat{f}_{\text{net}}(t, \mathbf{x}, \theta_{\text{net}})$. Hence, the formulation of UDEs allows us to incorporate arbitrary levels of mechanistic knowledge. Here, θ_{net} are the weights and biases of the neural network and θ_{mech} are the interpretable parameters of the mechanistic equation. Considering all parameters, we define $\theta = (\theta_{\text{mech}}, \theta_{\text{net}}, \theta_{np})$ for scenarios in which the initial condition \mathbf{x}_0 is known. The parameters are jointly estimated from data.

(c) Sources of uncertainty

In general, we can (at least) formally identify two distinct types of uncertainty: *aleatoric* and *epistemic* uncertainty [9]. As they can guide the evaluation of model performance and its potential application to real-life scenarios, precise quantification of these types of uncertainty is essential. The aleatoric (statistical) uncertainty $\text{Var}(\mathbf{y}(t))$ is based on inherent random effects and, hence, irreducible. By introducing a noise model, we aim to describe the aleatoric uncertainty. Epistemic (systematic) uncertainty stems from a lack of knowledge and potential model misspecifications. The bias-variance decomposition of the mean squared prediction error illustrates these different types of uncertainties [26]:

$$\begin{aligned} \mathbb{E}_{y(t)}[\mathbb{E}_{\mathcal{D}}[(\mathbf{y}(t) - \hat{\mathbf{y}}(t))^2]] &= (\mathbb{E}_{\mathcal{D}}[\hat{\mathbf{y}}(t)] - \bar{\mathbf{y}}(t))^2 + \text{Var}_{\mathcal{D}}(\hat{\mathbf{y}}(t)) + \text{Var}(\mathbf{y}(t)) \\ &= \text{bias}^2 + \text{Var}_{\mathcal{D}}(\hat{\mathbf{y}}(t)) + \text{Var}(\mathbf{y}(t)). \end{aligned} \quad (2.3)$$

Hence, epistemic uncertainty can be decomposed further into model bias $(\mathbb{E}_{\mathcal{D}}[\hat{\mathbf{y}}(t)] - \bar{\mathbf{y}}(t))$ and variance $\text{Var}_{\mathcal{D}}(\hat{\mathbf{y}}(t))$. As described in the previous section, we generally do not know f and θ . Uncertainty in the estimates $\hat{\theta}$ (model estimation) and \hat{f} (model form) are sources of epistemic uncertainty. There exist various methods for the estimation of epistemic uncertainty, as discussed in §3. However, the model bias is often neglected by assuming $\mathbb{E}_{\mathcal{D}}[\hat{\mathbf{y}}(t)] = \bar{\mathbf{y}}(t)$ and reducing the epistemic uncertainty to approximation uncertainty. While we will follow this assumption, we cannot guarantee a negligible model uncertainty: SciML is typically applied to the low to medium data regime [27]. Although neural networks are universal approximators, making them asymptotically unbiased, a bias is typically still observed in the low to medium data regime [28,29].

UDEs are located at the interface of neural networks and mechanistic dynamical modelling. While regularization is vital for neural networks [30], so is the exhaustive exploration of the parameter space for mechanistic models where one is interested in a global solution. It is not trivial to find the right balance between these, which is one of the reasons why parameter uncertainty, i.e. estimation uncertainty, is of quite some importance for UDEs. Furthermore, the numerical precision of the ODE solver and data sparsity may influence the quality of parameter estimation.

3. Methodology for epistemic uncertainty quantification of UDEs

(a) General setting

In this study, we will explore epistemic uncertainty arising as a result of parameter uncertainty, keeping the model form f fixed per problem setting. Bayes' rule provides a formulation for this

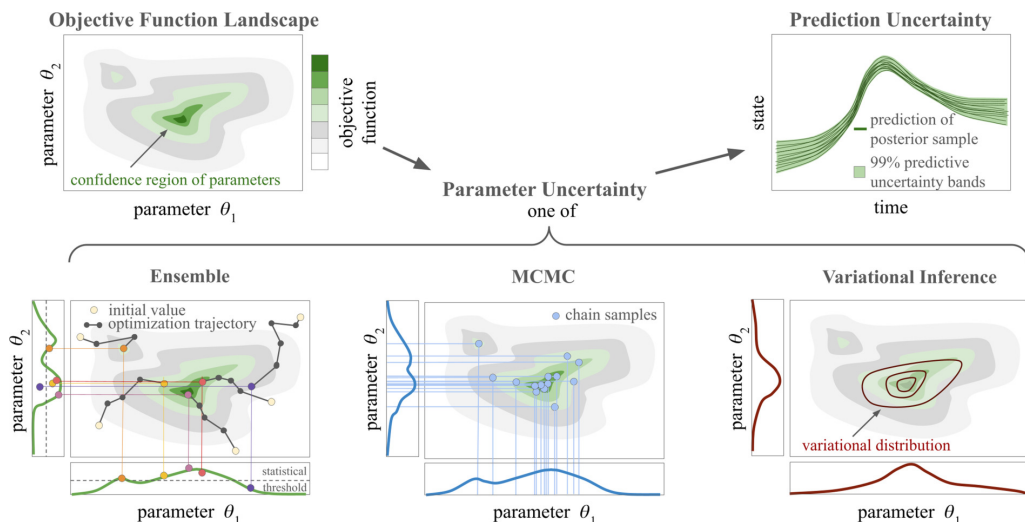


Figure 1. Overview of the presented uncertainty quantification methods. Given an unknown objective function landscape, we estimate the posterior distribution of the parameters using UDE ensembles, MCMC methods or Variational Inference. Based on random draws from the posterior distribution, predictions of the state trajectory are made, yielding lower and upper bounds for a 99% prediction interval.

uncertainty. The posterior density $p(\theta|\mathcal{D})$ can be described in terms of the likelihood $p(\mathcal{D}|\theta)$ and prior $p(\theta)$:

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{\int p(\mathcal{D}|\theta)p(\theta)d\theta}. \quad (3.1)$$

From this, the posterior predictive distribution is obtained as:

$$p(y(t)|\mathcal{D}) = \int p(y(t)|\theta)p(\theta|\mathcal{D})d\theta. \quad (3.2)$$

The prior distribution for the mechanistic parameters is usually chosen based on knowledge about the underlying processes. For the neural network parameters the choice of the prior distribution is less trivial. Commonly, an isotropic Gaussian prior is chosen [31]. Recently, it has been shown that especially for deep and flexible neural networks, this can cause drawbacks like the cold-posterior effect [32]. Specifying the correct prior is still a highly investigated research topic, and several options are discussed as alternatives for isotropic Gaussian priors [33]. One comparatively simple option is a Gaussian prior with a non-diagonal covariance matrix, allowing for correlation between different parameters [31].

(b) Uncertainty quantification methods

In the design phase for the work, we considered a broad spectrum of epistemic UQ methods that are theoretically suitable for dynamical systems such as UDEs. This included UQ methods based on multistart ensemble, MCMC sampling, Variational Inference, PLs/Posteriors and the FIM. In light of aspects related to scalability and applicability (e.g. due to identifiability constraints), we decided to disregard PL/Posterior calculation as well as methods using the FIM. For a detailed discussion, we refer to appendix A. The UQ methods for this study are illustrated in figure 1, and a description is provided in the following.

(i) Ensemble-based uncertainty quantification

To construct an ensemble-based UQ method, we consider combining ideas from deep learning [13] and mechanistic dynamical modelling [8] to define ensemble members that are customized to UDEs. After defining a prior distribution, encoding mechanistic knowledge when possible, we sample m initial values $\Theta_{\text{init}} = \{\theta_{\text{init}}^1, \dots, \theta_{\text{init}}^m\}$ from it. Then, m UDEs are trained, each starting with one element of Θ_{init} . The resulting optima are denoted by $\Theta = \{\hat{\theta}^1, \dots, \hat{\theta}^m\}$. During the optimization procedure, two issues have to be solved: overfitting and non-optimal local minima. To overcome overfitting, we use early stopping (in combination with an L2 regularization of the neural network parameters). Note that the necessary train-validation split yields a second source of randomness in our ensemble implementation. Additionally, like many dynamical approaches, UDEs can face convergence issues and numerical instabilities, which results in estimates with low likelihood. To address the issue of ending the optimization at non-optimal local minima, only a subset of the resulting estimators in Θ are accepted as ensemble members. We assume that m is large enough, such that the estimated maximum-likelihood estimate over all elements in Θ , $\hat{\theta}_{\text{MLE}}$, is approximately equal to the theoretical maximum-likelihood estimate θ_{MLE} . Based on a likelihood-ratio test, we evaluate whether the likelihoods of the other parameters in Θ significantly differ from $\hat{\theta}_{\text{MLE}}$ (a method commonly used in Systems Biology [8,16]). Hence, we keep parameter values which achieve a likelihood similar to the best estimate. The test statistic is defined as

$$\lambda(\hat{\theta}) = -2(\log(p(\mathcal{D}|\theta_{\text{MLE}})) - \log(p(\mathcal{D}|\hat{\theta}))) \approx -2(\log(p(\mathcal{D}|\hat{\theta}_{\text{MLE}})) - \log(p(\mathcal{D}|\hat{\theta}))) \quad (3.3)$$

and evaluated for all $\hat{\theta} \in \Theta$. The threshold for $\lambda(\hat{\theta})$ is given by the α -quantiles of a χ^2 distribution with n_f degrees of freedom, where $n_f = 1$ can provide a lower bound on the uncertainty. The parameters which meet the criterion provide an ensemble, which can be used for parameter and prediction uncertainty analysis (see [8]).

(ii) MCMC-based uncertainty quantification

Bayesian models perform Bayesian inference for the parameters of a model based on equation (3.1). Usually, the posterior has no closed-form solution. Instead, one relies on approximate inference algorithms like Variational Inference or MCMC-based methods [31].

MCMC-based UQ exploits parameter samples of the posterior distribution. These samples can be obtained using MCMC sampling approaches such as (adaptive) Metropolis–Hasting, parallel tempering and Hamiltonian Monte Carlo (HMC) algorithms. HMC algorithms leverage gradient information to search the parameter space efficiently, which has proven to be useful in dynamical modelling [34] and Bayesian neural networks [31]. To mitigate the problem that plain HMC is highly sensitive to parameters of the algorithm, one can use the HMC extension No-U-Turn Sampler (NUTS) [35]. At the moment, the NUTS algorithm is state-of-the-art for Bayesian neural networks [36], showing the best performance in the context of neural ODEs [22]. Yet, even with these approaches, the sampler often struggles to explore more than one mode. Parallel tempering algorithms operate with chains on different kinetic energy levels (temperatures). While low-temperature chains explore individual modes, high-temperature chains more easily traverse through the parameter space. By swapping the states of different chains under certain circumstances, parallel tempering algorithms aim to explore multimodal distributions more efficiently [37].

(iii) Variational Inference-based uncertainty quantification

Variational Inference approximates the posterior distribution $p(\theta|\mathcal{D})$ using a parametric distribution $q(\theta|\psi)$ with distribution parameters $\psi \in \mathbb{R}^{n_\psi}$ [23]. The Kullback–Leibler divergence is commonly used as an objective function, describing the discrepancy between the two distributions. This reduces the inference problem to defining an appropriate variational distribution $q(\theta|\psi)$ and estimating its parameters ψ . Commonly used options are, for example, a Gaussian

distribution $q(\theta|\psi) = \mathcal{N}(\theta|\mu, \Sigma)$, if no parameter bounds are given, or a scaled beta distribution $q(\theta|\psi) = c \cdot \text{Beta}(\theta|a, b)$ if $\theta \in (0, c)$.

4. Performance evaluation of methods: insights from synthetic examples

To assess the performance of the aforementioned UQ methods, we performed experiments on several synthetic problems. By using synthetic problems, we can evaluate the methods' performance by comparing their results to the known data-generating process.

(a) Model formulation

In this study, we considered three different models (SEIR Pulse, SEIR Waves and Quadratic Dynamics) with two different noise distributions (Gaussian and negative Binomial). For each noise distribution, we investigated several noise parameter settings. Electronic supplementary material, table S1 and appendix C provide an overview of the considered models and scenarios.

In the main part of the paper, we present the results based on the SEIR model, a compartmental model describing the dynamics of infectious diseases. This model describes the number of susceptible (S), exposed (E), infectious (I) and recovered (R) individuals using a system of differential equations:

$$\frac{dS}{dt} = -\beta(t) \frac{SI}{N}, \quad \frac{dE}{dt} = \beta(t) \frac{SI}{N} - \alpha E, \quad \frac{dI}{dt} = \alpha E - \gamma I, \quad \frac{dR}{dt} = \gamma I, \quad (4.1)$$

where β is the transmission rate, α the transition rate, γ the recovery rate and $N = S + E + I + R$ the population size.

We create synthetic data for two scenarios, differing the time-dependent transmission rate $\beta(t)$ (see electronic supplementary material, figure S1):

(1) In the **SEIR Pulse** scenario, we define the transmission rate as

$$\beta(t) = \begin{cases} 0.05 & \text{if } 15 < t < 30, \\ 0.5 & \text{else,} \end{cases} \quad (4.2)$$

which emulates a time-restricted political intervention.

(2) In the **SEIR Waves** scenario, we define the transmission rate as

$$\beta(t) = 0.3 \cdot \cos\left((-1 + \sqrt{1 + 4t}) \cdot 1.5 + 0.25 \cdot \pi\right) + 0.4, \quad (4.3)$$

which is an oscillating function for emulating the potentially complex time-dependence of contact behaviour in combination with potential virus evolution.

For the synthetic data generation, we assumed that the state variables I and R are observed at 30 time points. The parameters and initial conditions used for simulation are provided in appendix B.

In practice, the basic transmission process of an infectious disease is often known, but the parameter values and the time-dependence of the transition rate are unknown. To capture this case, we assumed the mechanistic formulation of the differential equation to be known, but the constant rate parameters α and γ as well as the time-dependent transmission rate β to be unknown. A neural network was used to model β , where we use a tanh-parameterization to ensure bounds and easily capture fast-changing dynamics. For the estimation of the noise parameter, a log transformation was used; and for α and γ , a tanh transformation was used (see appendix D).

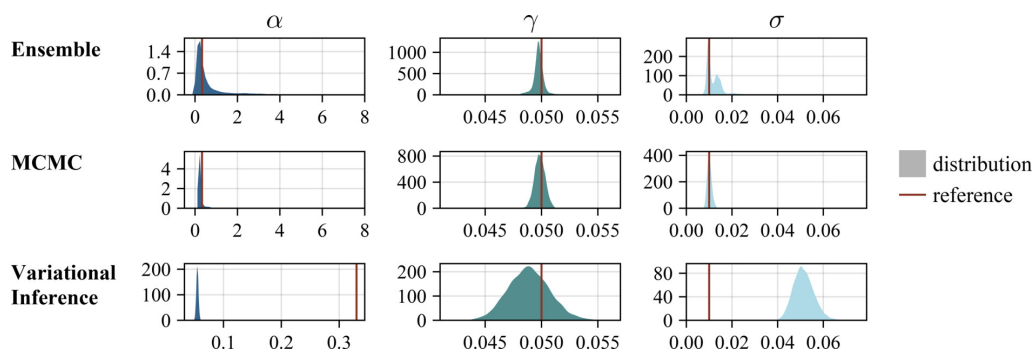


Figure 2. Comparison of the UQ methods on the SEIR Waves problem (Gaussian noise, $\sigma = 0.01$), visualization of estimated parameter values. The visualization for the time-dependent parameter $\beta(t)$ is provided in appendix I.

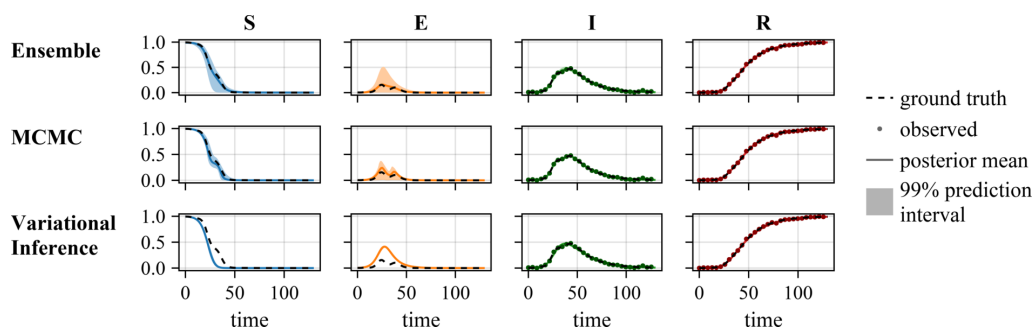


Figure 3. Comparison of the UQ methods on the SEIR Waves problem (Gaussian noise, $\sigma = 0.01$), visualization of estimated trajectories.

(b) Results

In the following, we present the results for ensemble-, MCMC- and Variational Inference-based UQ. Implementation details such as the prior definition and hyperparameter settings are provided in appendix D.

(i) Ensemble-based uncertainty quantification

The assessment of the results for ensemble-based UQ revealed overall a good performance: The true parameters are contained in the uncertainty intervals (figure 2) and simulated trajectories of the dynamics of the model using the true parameter values are contained in the prediction intervals (figure 3). Indeed, also the (unknown) standard deviation used for synthetic data generation are contained in the corresponding uncertainty intervals (figure 2). Yet, the long tail of the distribution indicates that the ensemble contains samples that overestimate the aleatoric uncertainty. Furthermore, while the underlying dynamics of the unobserved states could be recovered, the UQ for the transmission rate β yields broad prediction intervals (electronic supplementary material, figure S12). A broad band of trajectories of β yields reasonable values for the observed states I and R . This is not unexpected since β influences the state change only for time points with $I \cdot S \gg 0$. We would expect a reliable reconstruction only in these regimes. Ensemble members with smaller negative log-likelihood values tend to show dynamics more closely related to the dynamics of the data generation process for these time points.

We observe that, as expected, the ensemble-based UQ method yields larger prediction uncertainty bounds with increasing aleatoric noise (electronic supplementary material, figure S13). Small fluctuations in the trajectory cannot be captured easily within a setting of negative binomial noise, as is indicated by the ensemble mean trajectory for I . In the SEIR problem scenarios, the role

of the neural network is well isolated from the other dynamical components. In the quadratics dynamics scenario, this is different, because the neural network can in principle completely replace known mechanistic dynamics and describe all the dynamics. Hence, the mechanistic part is only a soft constraint on the form of the whole dynamics. A consequence of this is visualized in electronic supplementary material, figure S14: the predictions of some ensemble members quickly deviate from the reference dynamics outside the data domain.

One difficulty of the ensemble-based UQ method was the choice of a reasonable threshold. As visualized in electronic supplementary material, figure S11, subselecting a fraction of the best-performing models (which is equivalent to a different significance level for the χ^2 -test) can result in widely different confidence bands. An exemplary waterfall plot displaying the selection of the ensemble members based on likelihood values is provided in electronic supplementary material, figure S9 and shows that there is no clear convergence to a minimum objective function value for UDEs. A major advantage of the ensemble-based UQ method is its flexible parallelisability: Every candidate member of the ensemble can be trained independently of one another. For 10 000 candidate ensemble members, the training took between 4–12 h using 20 CPU cores.

(ii) MCMC-based uncertainty quantification

We implemented Bayesian UDEs using a NUTS and parallel tempering sampling algorithm to compare different and potentially suitable algorithms for UDEs. When sampling, the biggest issue of overparameterized models is the exploration of multiple modes. Neural networks specifically tend to have by construction various symmetries in the loss landscape [38], resulting in the possibility that no additional predictive information is added even if multiple modes are explored. We systematically experimented with different numbers of chains and samples. However, similar to what is observed with neural networks in classical supervised learning tasks [36], MCMC chains neither mix well nor properly converge in the context of UDEs. Our analysis of the sample distributions indicated that parallel tempering algorithms achieved a better exploration than NUTS. Furthermore, the initialization of the sampling at optimization endpoints—a common recommendation for dynamical systems [39]—was beneficial for UDEs.

MCMC-based UQ using NUTS and parallel tempering was computationally much more demanding than ensemble-based UQ. The generation of approximately 10 000 samples using parallel tempering required 5–7 days on a CPU node with 20 CPU cores. Yet, the resulting Bayesian confidence intervals possess a clear interpretation and there are no additional hyperparameters. Indeed, while challenging to apply, the parallel tempering runs provided posterior samples, which covered the true parameter values (figure 2). Furthermore, the corresponding prediction intervals contained the true dynamics of the process (figure 3).

(iii) Variational Inference-based uncertainty quantification

We performed Variational Inference using a mean-field approximation of the posterior with multivariate normal base distributions. This is a common choice for a broad range of uncertainty analysis problems [10]. Yet, the choice of approximation and base distribution can have a substantial impact on the approximation quality.

Variational Inference-based UQ required <24 h on a single CPU core and was therefore computationally less demanding than ensemble- and MCMC-based UQ. However, the assessment of the results revealed that Variational Inference provided biased estimates. The approximated posterior distribution did not cover the true values of α and σ (figure 2). Furthermore, while the observations are fitted, Variational Inference underestimates the uncertainty on the trajectories of the unobserved state variables (figure 3).

(iv) Comparison of methods

The three considered methods tackle the problem of UQ from different angles. This results in differences concerning data usage and incorporation of prior knowledge, as well as

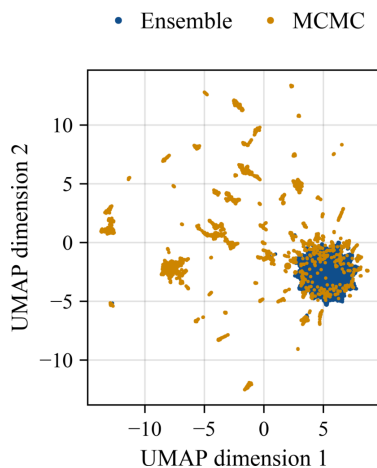


Figure 4. UMAP for the set of parameters obtained by the MCMC and ensemble-based method. The parameter vectors contained mechanistic and neural network parameters.

posterior coverage. Since the ensemble-based UQ method is based on the optimization of an over-parametrized model, a train-validation split is necessary to implement early stopping and avoid overfitting on the training data. The combined dataset is only used for subselecting from ensemble candidate models. For MCMC- and Variational Inference-based methods, the whole dataset is used in every step of the algorithms. While the incorporation of prior knowledge in the dynamic equations, noise model and observable mapping is independent of the UQ method, assumptions about parameter values are treated differently. For the ensemble-based method, the parameter prior influences the start points of the optimization process. Afterwards, we only encode upper and lower bounds that restrict the parameter update steps. For both MCMC- and Variational Inference-based UQ, the prior distributions influence the posterior in every update step of the algorithms.

Our comparison of UQ results across methods revealed that ensemble- and MCMC-based methods provide assessments/predictions, while the standard Variational Inference method provides biased estimates and is unable to capture the dynamics of the unobserved state variables. Interestingly, while one might expect that this changes for small noise levels, our findings suggest that it holds independent of measurement noise (appendix F), which might be due to the over-parametrization in the UDE and corresponding non-identifiabilities.

Yet, we also identified substantial differences between ensemble- and MCMC-based UQ method results. A dimension reduction of the parameter samples shows different characteristics. The uniform manifold approximation and projection (UMAP) [40] of the collection of all samples using parameter values as features reveals that: (i) ensembles derived from optimization results cluster; and (ii) parallel tempering yields a more patchy pattern and several additional clusters (figure 4). While on first glance this suggests that parallel tempering achieves a better exploration, it also seems to miss high-quality parameter vectors identified using optimization. To assess this further, we studied the model predictions for an initial condition for which no measurements were available (figure 5). The ensemble-based UQ yields substantially broader intervals than the MCMC-based UQ. Predictions of the MCMC-based UQ still cover the model simulation for the true data.

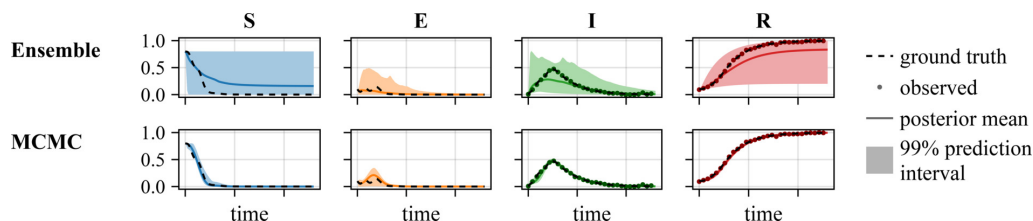


Figure 5. Comparison of prediction performance using the same setting as in figure 3 but with a new initial condition $\mathbf{x}_0 = (0.8, 0.1, 0.0, 0.1)$.

5. Conclusions and future perspectives

UDEs model the dynamics of observed and unobserved states using a combination of interpretable mechanistic and ANN parameters. UQ for both mechanistic parameters and state trajectories is critically important, especially in fields like healthcare and epidemiology. However, performing UQ for UDEs poses unique challenges, primarily due to the high dimensionality of the parameter space and the complex structure of the loss or posterior function landscape.

Our assessment of UQ methods revealed that ensemble- and MCMC-based methods perform better than Variational Inference. This observation holds true across different noise levels. Nevertheless, defining a suitable threshold for ensemble-based methods requires further investigation, and the convergence of MCMC methods remains problematic.

Our findings underscore the need for further methodological developments. Building upon the work of [36], developing methods for automatic symmetry removal could improve computational efficiency in UQ for UDEs and enable scaling to larger problems. Furthermore, exploring hybrid UQ approaches that combine the resource efficiency of ensemble methods with the statistical interpretability of MCMC-based methods might be advantageous. Lastly, investigating model structure uncertainty may provide valuable insights into how mechanistic terms are absorbed by the neural network.

The results of our study likely extend beyond UDEs and may also apply to alternative machine learning-based modelling approaches. An assessment of ANNs designed to control dynamical systems [41] would be of particular interest. Furthermore, extending this work to stochastic models, particularly neural stochastic differential equations [42], would be beneficial. Accordingly, while this study provides novel insights, it can only serve as a starting point for further research.

Data accessibility. Code to run the presented experiments and regenerate simulation data can be found in the Github repository [43].

Supplementary material is available online [44].

Declaration of AI use. We have not used AI-assisted technologies in creating this article.

Authors' contributions. N.S.: conceptualization, formal analysis, software, visualization, writing—original draft; D.FdP.: conceptualization, writing—review and editing; W.W.: conceptualization, funding acquisition, writing—review and editing; J.H.: conceptualization, funding acquisition, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Funding. This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy (EXC 2047—390685813, EXC 2151—390873048), by the German Federal Ministry of Education and Research (BMBF) (INSiDe- grant number 031L0297A), the ERC Consolidator Grant (INTEGRATE- grant agreement number 101126146), and by the University of Bonn (via the Schlegel Professorship of J.H.). W.W. received funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" Programme.

1. Rico-martínez R, Krischer K, Kevrekidis IG, Kube MC, Hudson JL. 1992 Discrete- vs. continuous time nonlinear signal processing of Cu electrodisolution data. *Chem. Eng. Commun.* **118**, 25–48. (doi:[10.1080/00986449208936084](https://doi.org/10.1080/00986449208936084))
2. Wang YJ, Lin CT. 1998 Runge-Kutta neural network for identification of dynamical systems in high accuracy. *IEEE Trans. Neural Netw.* **9**, 294–307. (doi:[10.1109/72.661124](https://doi.org/10.1109/72.661124))
3. Chen RTQ, Rubanova Y, Bettencourt J, Duvenaud D. 2018 Neural ordinary differential equations. In *NeurIPS 2018* (S Bengio, H Wallach, H Larochelle, K Grauman, N Cesa-Bianchi, R Garnett eds), pp. 6572–6583. Red Hook, NJ: Curran Associates.
4. Baker N, Alexander F, Bremer T, Hagberg A, Kevrekidis Y, Najm H, Parashar M. *Workshop report on basic research needs for scientific machine learning: core technologies for artificial intelligence*. (doi:[10.2172/1478744](https://doi.org/10.2172/1478744))
5. Faroughi SA, Pawar NM, Fernandes C, Raissi M, Das S, Kalantari NK, Kourosh Mahjour S. 2024 Physics-guided, physics-informed, and physics-encoded neural networks and operators in scientific computing: fluid and solid mechanics. *J. Comput. Inf. Sci. Eng.* **24**, 040802. (doi:[10.1115/1.4064449](https://doi.org/10.1115/1.4064449))
6. Rackauckas C, Ma Y, Martensen J, Warner C, Zubov K, Supekar R, Skinner D, Ramadhan A, Edelman A. 2021 Universal differential equations for scientific machine learning. *Arxiv*. (doi:[10.21203/rs.3.rs-55125/v1](https://doi.org/10.21203/rs.3.rs-55125/v1))
7. Philipps M, Körner A, Vanhoefer J, Pathirana D, Hasenauer J. 2024 Non-negative universal differential equations with applications in systems biology. *IFAC-PapersOnLine* **58**, 25–30. (doi:[10.1016/j.ifacol.2024.10.005](https://doi.org/10.1016/j.ifacol.2024.10.005))
8. Villaverde AF, Raimundez E, Hasenauer J, Banga JR. 2023 Assessment of prediction uncertainty quantification methods in systems biology. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **20**, 1725–1736. (doi:[10.1109/TCBB.2022.3213914](https://doi.org/10.1109/TCBB.2022.3213914))
9. Hüllermeier E, Waegeman W. 2021 Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Mach. Learn.* **110**, 457–506. (doi:[10.1007/s10994-021-05946-3](https://doi.org/10.1007/s10994-021-05946-3))
10. Abdar M *et al.* 2021 A review of uncertainty quantification in deep learning: techniques, applications and challenges. *Inf. Fusion* **76**, 243–297. (doi:[10.1016/j.inffus.2021.05.008](https://doi.org/10.1016/j.inffus.2021.05.008))
11. Metropolis N, Ulam S. 1949 The Monte Carlo method. *J. Am. Stat. Assoc.* **44**, 335–341. (doi:[10.1080/01621459.1949.10483310](https://doi.org/10.1080/01621459.1949.10483310))
12. Jaakkola TS, Jordan MI. 2000 Bayesian parameter estimation via variational methods. *Stat. Comput.* **10**, 25–37. (doi:[10.1023/A:1008932416310](https://doi.org/10.1023/A:1008932416310))
13. Lakshminarayanan B, Pritzel A, Blundell C. 2017 Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems* (I Guyon, U Von Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, R Garnett eds), pp. 6405–6416. Red Hook, NJ: Curran Associates. <https://dl.acm.org/doi/10.5555/3295222.3295387>.
14. Villaverde AF, Bongard S, Mauch K, Müller D, Balsa-Canto E, Schmid J, Banga JR. 2015 A consensus approach for estimating the predictive accuracy of dynamic models in biology. *Comput. Methods Programs Biomed.* **119**, 17–28. (doi:[10.1016/j.cmpb.2015.02.001](https://doi.org/10.1016/j.cmpb.2015.02.001))
15. Gal Y, Ghahramani Z. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In *In PMLR*, vol. 48, pp. 1050–1059, <https://proceedings.mlr.press/v48/gal16.html>.
16. Kreutz C, Raue A, Kaschek D, Timmer J. 2013 Profile likelihood in systems biology. *FEBS J.* **280**, 2564–2571. (doi:[10.1111/febs.12276](https://doi.org/10.1111/febs.12276))
17. Jacquez JA, Greif P. 1985 Numerical parameter identifiability and estimability: integrating identifiability, estimability, and optimal sampling design. *Math. Biosci.* **77**, 201–227. (doi:[10.1016/0025-5564\(85\)90098-7](https://doi.org/10.1016/0025-5564(85)90098-7))
18. Raissi M, Perdikaris P, Karniadakis GE. 2019 Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.* **378**, 686–707. (doi:[10.1016/j.jcp.2018.10.045](https://doi.org/10.1016/j.jcp.2018.10.045))

19. Psaros AF, Meng X, Zou Z, Guo L, Karniadakis GE. 2023 Uncertainty quantification in scientific machine learning: methods, metrics, and comparisons. *J. Comput. Phys.* **477**, 111902. (doi:10.1016/j.jcp.2022.111902)
20. Akhare D, Luo T, Wang JX. DiffHybrid-UQ: uncertainty quantification for differentiable hybrid neural modeling. See <https://arxiv.org/pdf/2401.00161>.
21. Acquesta ECS, Portone T, Dandekar R, Rackauckas C, Bandy R, Huerta G, Dytzel I. 2022 *Model-form epistemic uncertainty quantification for modeling with differential equations: application to epidemiology*. (doi:10.2172/1888443)
22. Dandekar R, Chung K, Dixit V, Tarek M, Garcia-Valadez A, Vemula KV, Rackauckas C. Bayesian neural ordinary differential equations. See <https://arxiv.org/pdf/2012.07244>.
23. Gawlikowski J *et al.* 2023 A survey of uncertainty in deep neural networks. *Artif. Intell. Rev.* **56**, 1513–1589. (doi:10.1007/s10462-023-10562-9)
24. Contento L *et al.* 2023 Integrative modelling of reported case numbers and seroprevalence reveals time-dependent test efficiency and infectious contacts. *Epidemics* **43**, 100681. (doi:10.1016/j.epidem.2023.100681)
25. Chan S, Chu J, Zhang Y, Nadarajah S. 2021 Count regression models for COVID-19. *Physica A* **563**, 125460. (doi:10.1016/j.physa.2020.125460)
26. Gruber C, Schenk PO, Schierholz M, Kreuter F, Kauermann G. Sources of uncertainty in machine learning: a statisticians' view. See <https://arxiv.org/pdf/2305.16703>.
27. Peng GCY *et al.* 2021 Multiscale modeling meets machine learning: what can we learn? *Arch. Comput. Methods Eng.* **28**, 1017–1037. (doi:10.1007/s11831-020-09405-5)
28. Belkin M, Hsu D, Ma S, Mandal S. 2019 Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proc. Natl Acad. Sci. USA* **116**, 15849–15854. (doi:10.1073/pnas.1903070116)
29. Neal B, Mittal S, Baratin A, Tantia V, Scicluna M, Lacoste-Julien S, Mitliagkas I. *A modern take on the bias-variance tradeoff in neural networks*. See <https://arxiv.org/abs/1810.08591>.
30. Goodfellow I, Bengio Y, Courville A. 2016 *Deep learning*. Cambridge, MA: MIT Press.
31. Murphy KP. 2023 *Probabilistic machine learning: advanced topics*. Cambridge, MA: MIT Press. See <http://probml.github.io/book2>.
32. Wenzel F *et al.* How good is the Bayes posterior in deep neural networks really? (eds HD III, A Singh). In *PMLR*, vol. 119, pp. 10248–10259. PMLR. <https://proceedings.mlr.press/v119/wenzel20a.html>.
33. Fortuin V. 2022 Priors in Bayesian deep learning: a review. *Int. Statistical Rev.* **90**, 563–591. (doi:10.1111/insr.12502)
34. Schälte Y *et al.* 2023 pyPESTO: a modular and scalable tool for parameter estimation for dynamic models. *Bioinformatics* **39**, btad711. (doi:10.1093/bioinformatics/btad711)
35. Hoffman MD, Gelman A. 2014 The no-u-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.* **15**, 1593–1623. <http://jmlr.org/papers/v15/hoffman14a.html>
36. Sommer E, Wimmer L, Papamarkou T, Bothmann L, Bischl B, Rügamer D. Connecting the dots: is mode-connectedness the key to feasible sample-based inference in Bayesian neural networks?. See <https://arxiv.org/pdf/2402.01484>.
37. Syed S, Bouchard-Côté A, Deligiannidis G, Doucet A. 2022 Non-reversible parallel tempering: a scalable highly parallel MCMC scheme. *J. R. Stat. Soc. B* **84**, 321–350. (doi:10.1111/rssb.12464)
38. Wiese JG, Wimmer L, Papamarkou T, Bischl B, Günnemann S, Rügamer D. 2023 Towards efficient MCMC sampling in Bayesian neural networks by exploiting symmetry. In *Machine learning and knowledge discovery in databases: research track*, pp. 459–474. Cham, Switzerland: Springer Nature Switzerland. (doi:10.1007/978-3-031-43412-9_27)
39. Hass H, Loos C, Raimúndez-Álvarez E, Timmer J, Hasenauer J, Kreutz C. 2019 Benchmark problems for dynamic modeling of intracellular processes. *Bioinformatics* **35**, 3073–3082. (doi:10.1093/bioinformatics/btz020)
40. McInnes L, Healy J, Melville J. 2020 UMAP: uniform manifold approximation and projection for dimension reduction. *J. Open Source Softw.* **3**, 861. (doi:10.21105/joss.00861)
41. Kidger P, Morrill J, Foster J, Lyons T. Neural controlled differential equations for irregular time series.

42. Jia J, Benson AR. Neural jump stochastic differential equations. See: <https://arxiv.org/abs/1905.10403>
43. Schmid N. 2024 *Code for 'assessment of uncertainty quantification in universal differential equations'*. See https://github.com/schminin/UQ_in_UDEs.
44. Schmid N, del Pozo F, Waegeman D, Hasenauer W. 2025 Supplementary material from: Assessment of uncertainty quantification in universal differential equations. Figshare. (doi:10.6084/m9.figshare.c.7718169)