Mini-batching ecological data to improve ecosystem models with machine learning

Victor Boussange[a,b,*], Pau Vilimelis Aceituno[c], and Loïc Pellissier[a,b]

[a]Swiss Federal Research Institute WSL, Birmensdorf, Switzerland

[b]Landscape Ecology, Institute of Terrestrial Ecosystems, Department of Environmental System Science, ETH Zürich, Zürich, Switzerland

[c]Institute of Neuroinformatics, ETH Zürich and University of Zürich, Zürich, Switzerland

[*]Corresponding author,

Email: bvictor@ethz.ch

July 25, 2022

Abstract

Ecosystems are involved in global biogeochemical cycles that regulate climate and provide essential services to human societies. Mechanistic models are required to describe ecosystem dynamics and anticipate their response to anthropogenic pressure, but their adoption has been limited in practice because of issues with parameter identification and because of model inaccuracies. While observations could be used to directly estimate parameters and improve models, model nonlinearities as well as shallow, incomplete and noisy datasets complicate this process. Here, we propose a machine learning (ML) framework relying on a mini-batch method combined with automatic differentiation and state-of-the-art optimizers. By splitting the data into mini-batches with a short time horizon, we show both analytically and numerically that the mini-batch method regularizes the learning problem. When combined with the proposed numerical implementation, the resulting ML framework can efficiently learn the parameter of complex dynamical models and is a workhorse for model selection. We evaluate the performance of the ML framework in recovering the dynamics of a simulated food-web. We show that it can efficiently learn from noisy, incomplete and independent time series, accurately estimating the model parameters and providing reliable short-term forecasts. We further show that the ML framework can provide statistical support for the true generating model among several candidates. In summary, the proposed ML framework can efficiently learn from data and elucidate mechanistic pathways to improve our understanding and predictions of ecosystem dynamics.

Scientific machine learning | Inverse modelling | Ecosystem modelling | Model selection | Food-webs | Mechanistic model parametrization | Data assimilation

Author summary

Ecosystem models which explicitly represent ecological mechanisms are required to forecast ecosystem responses to global changes, but large mismatches with observations limit their predictive ability. To help address this major problem, we propose a novel machine learning (ML) method aiming at improving ecosystem models with data. The ML method is based on a learning strategy where the model is matched against small chunks of data, called mini-batches, and it involves numerical techniques commonly used in the training of neural networks. By benchmarking the performance of the ML method with a challenging food-web model, we show that our approach is robust against noise and partial observations, can process and combine the information contained in independent datasets, and can provide statistical support for the most adequate model among several candidates. Our proposed method therefore accommodates the reality of ecological datasets and our partial knowledge of ecosystem processes. By efficiently blending data and ecological theory with state-of-the-art ML techniques, our work offers novel tools to improve our understanding and predictions of ecosystem dynamics.

Contents

1 Introduction

Ecosystems are complex systems involving many interacting functional entities which together play a major role in regulating global biogeochemical cycles [1] and delivering essential services to humans [2]. Ecosystems currently face intense disruption from anthropogenic pressure, through pollution and land use [3, 4], and from climate change [5]. In order to anticipate the responses of ecosystems to these disruptions, models that can extrapolate ecological dynamics beyond observations are required [6]. A major challenge is that the processes driving ecological dynamics are nonlinear, resulting in complex responses and feedbacks [7]. Nonlinearity greatly affects the capacity of modelling approaches that do not incorporate specific biological knowledge to reliably project current trends into the future [8]. For instance, while methods based on statistical descriptions [9] and nonparametric methods [10, 11, 12] have adequate interpolation capabilites, they are ill-suited for extrapolating beyond observed trends [8, 13]. In contrast, mechanistic ecosystem models integrate constraints on the expected dynamics by explicitly modelling interactions, feedback loops and dependencies between ecosystem components [14]. While this should ensure a more robust forecast under large disruptions [15], ecosystems models suffer in practice from parametrization issues, i.e. inaccuracies in the mathematical formulation of the processes and issues in identifying the correct parameter values [16]. These drawbacks have limited their broad adoption [13]. Learning the parametrization of ecosystem models from observation data by blending specific biological knowledge and ML methods could improve our representation of ecosystem processes and help us to anticipate ecosystem responses to global changes.

The parametrization of ecosystem models can be indirectly learnt from observations by calibrating the parameters from the data collectively using inference methods. These methods proceed by maximizing the posterior probability of the parameters given the observations, but their success is subject to a number of issues, some of which specifically relating to ecosystem model properties. Among these issues, the exploration of the posterior landscape demands repeated model simulations, but ecosystem models are usually associated with a high computational cost that limits the number of possible runs [17]. Additionally, the complexity of processes requires a large number of parameters. Due to the curse of dimensionality [6], this complicates the exploration of the posterior distribution and can further leave many parameters poorly constrained [18]. The limited availability of observations, which are usually composed of multiple partial short-term time series [19, 20], accentuates the lack of parameter constraints. Moreover, ecosystem dynamics can be strongly dependent on the initial conditions (ICs) and show a chaotic behaviour [21, 22, 23], or can be associated with a large panel of dynamics depending on the parameter values. In such case, small perturbations of the ICs or parameter values

lead to large divergences in the model outcomes, causing numerical problems in finding the most probable parameters [16]. Last but not least, in contrast to fields such as climate and weather modelling, the derivation of fundamental processes regulating ecosystems is far from being established [24, 20, 25], resulting in inaccurate mechanistic pathways and uncertain mathematical formulations [26], limiting extrapolation to unseen data. To summarize, the parametrization of ecosystem models requires inference methods that are robust despite the models' complexity, the limited observation data, and the inaccurate description of ecological processes.

A variety of data assimilation and ML methods are increasingly being used to parametrize ecosystem models. Bayesian inference with Markov Chain Monte Carlo methods, used in [27, 28, 29, 30, 31], offer the advantage of quantifying uncertainties by inferring the full posterior probability distribution of the unknown parameters. This is achieved by a global exploration of the parameter space, which makes Bayesian methods computationally expensive and particularly prone to the curse of dimensionality [32]. Simulated annealing [33], genetic algorithms [34], and sequential methods such as extended Kalman filtering and ensemble Kalman methods [35, 36, 37] have been used as alternatives, but are similarly subject to the curse of dimensionality and demand a large number of model evaluations. Variational methods rely on the model adjoint, i.e. the model sensitivity to the parameters, to explore more efficiently the parameter space, iteratively updating the parameter estimates using the gradient of the posterior landscape. Such methods therefore demand less evaluation [20], which explains their wide adoption in the field of artificial intelligence to train highly parametrized neural networks (up to the order of 10^8 parameters [38]) and their use in calibrating marine ecosystem models [39, 40, 41, 42] (see [20] for a review) and terrestrial ecosystem models [43, 44, 45]. However, as the complex dynamics of ecosystem models tend to be associated with rugged posterior landscapes, variational methods are prone to converging to local minima, making variational methods very sensitive to the choice of initial model parameters [46, 20]. Ecosystem models are specified as differential equations that depend not only on parameters but also on ICs. The state-dependency of ecosystems means that neglecting the estimation of initial ICs might compromise the correct fitting of the parameters and the forecast skill [27]. However, few of the aforementioned studies have addressed the problem of IC estimation (but see [42]). Finally, the numerical implementation of variational methods is also challenging, as the model adjoint is difficult to obtain and maintain as the model is modified [47, 42, 37]. Novel methods for model parametrization are emerging, thanks to advances in the field of artificial intelligence [48, 49, 50, 51], providing new opportunities to better address these issues.

Here, we propose a ML framework relying on a mini-batch method inspired by multiple shooting methods [52] and on automatic differentiation and state-of-the-art optimizers to efficiently learn the parametrization of ecosystem models from observation data. The mini-batch method divides the training problem into mini-batches

with a short time horizon. We show analytically how this learning strategy regularizes the ill-behaviour of the loss function arising from the strong nonlinearities of ecosystem models. We implement the mini-batch method in the software ecosystem SciML [53], which provides advanced optimizers and allows the automatic generation of efficient and accurate model adjoints, leading to excellent performance. The resulting ML framework makes it possible to efficiently combine the information contained in short, independent time series, and is a workhorse for performing model selection and improving model accuracy. We evaluate the performance of the ML framework in recovering the chaotic dynamics of simulated food webs. We assume a perfect-model setting and test the capacity of the ML framework to recover the true parameters and provide forecasts based on noisy and incomplete observations, and we explore its efficiency in combining the information from multiple time series. Additionally, we investigate whether the ML framework can recover the most appropriate model structure among candidate models. By blending biological knowledge and ML methods, the proposed ML framework is interpretable and data-efficient, and it facilitates mechanism discovery. The proposed approach thus shows promise in improving our ability to understand and forecast ecosystem dynamics.

2 Machine learning framework for ecosystem models

2.1 Ecosystem model parametrization as a learning problem

2.1.1 Ecosystem models

Ecosystem models generally consist of a system of ordinary differential equations (ODEs) of the form

$$\begin{aligned} \dot{x}(t) &= f(t, x(t), p) \\ x(0) &= x_0 \\ y(t) &= h(x(t)) + \epsilon(t) \end{aligned} \tag{1}$$

where $x(t) \in \mathbb{R}^m$ is a vector of state variables that might represent species abundance, resources availability or functional group biomass, $y(t) \in \mathbb{R}^d$ is a vector of observables that contains a subset or aggregates of the state variables, and $p \in \mathbb{R}^q$ is the model parameter vector. h is a function that maps the state variables to the observables, and we assume that the observables are contaminated with a white noise ϵ of Gaussian type, with zero mean and variance–covariance matrix Σ_y. Denoting by $\theta = (x_0, p)$ the vector containing the ICs and the parameters, the model may be viewed as a map $\mathcal{M}$ parametrized by time t that takes the parameters θ to the

state variables x

$$\begin{aligned}\mathcal{M}(t,\theta) &= x(t) \\ &= \int_0^t f(s,x(s),p)ds + x_0\end{aligned} \tag{2}$$

2.1.2 Inverse modelling

Taking expectations over the noise realizations yields $\mathbb{E}[y(t)] = h(\mathcal{M}(t,\theta))$, and it follows that the conditional likelihood of each observation $y_k \equiv y(t_k)$, given the parameters θ and the model $\mathcal{M}$ denoted by $p(y_k|\theta,\mathcal{M})$, follows the distribution of the residuals $\epsilon_k \equiv \epsilon(t_k) = y(t_k) - h\left(\mathcal{M}(t_k,\theta)\right)$, which corresponds to the multivariate normal distribution $\mathcal{N}_{0,\Sigma_y}$. Following a Bayesian approach, the parametrization of the ecosystem model can be performed on the basis of the parameter and model posterior probability $p(\theta,\mathcal{M}|\mathbf{y}_{1:K})$, i.e. the conditional probability density of the parameter values θ and the model $\mathcal{M}$ given the data, given by

$$p(\theta,\mathcal{M}|\mathbf{y}_{1:K}) \propto p(\mathbf{y}_{1:K}|\theta,\mathcal{M})p(\theta,\mathcal{M}) \tag{3}$$

where $\mathbf{y}_{1:K} = (y_1,\ldots,y_K)$, $p(\mathbf{y}_{1:K}|\theta,\mathcal{M})$ is the product of the conditional likelihood of each observation y_k

$$\begin{aligned}p(\mathbf{y}_{1:K}|\theta,\mathcal{M}) &= \prod_{i=1}^{K} p(y_i|\theta,\mathcal{M}) \\ &= \prod_{k=1}^{K} \frac{1}{\sqrt{(2\pi)^d|\Sigma_y|}} \exp\left(-\frac{1}{2}\epsilon_k^T\Sigma_y^{-1}\epsilon_k\right)\end{aligned} \tag{4}$$

and $p(\theta,\mathcal{M})$ is the prior distribution of the model and its associated parameter values. The model $\mathcal{M}$ is included in the probabilistic quantities in order to accommodate multiple candidate models (see Section 3.3).

A variational method to obtain a Bayesian estimate of θ involves maximizing $p(\theta,\mathcal{M}|\mathbf{y}_{1:k})$ to obtain the maximum a posteriori (MAP) estimator [54], which is equivalent to a maximum likelihood approach under a uniform prior distribution of the parameters, i.e. when no prior information on the parameter values is used [20]. Observing that maximizing $p(\theta,\mathcal{M}|\mathbf{y}_{1:K})$ is equivalent to minimizing $-\log p(\theta|\mathbf{y}_{1:K},\mathcal{M})$ and assuming a normal prior distribution of the parameters $\mathcal{N}_{p_b,\Sigma_p}$, one can obtain the MAP $\hat{\theta}$

$$\hat{\theta} = \arg\min_{\theta} L_{\mathcal{M}}(\theta) \tag{5}$$

where

$$L_{\mathcal{M}}(\theta) = \frac{1}{2}\left[\sum_{k=1}^{K-1} \|y_k - h\left(\mathcal{M}(t_k,\theta)\right)\|^2_{\Sigma_y} + \|p - p_b\|^2_{\Sigma_p}\right] \tag{6}$$

[55, 56] and where we use the notation $\|y\|^2_\Sigma = y\Sigma^{-1}y^{\mathbf{T}}$. Eq. (6) is similar to a traditional least squares function commonly used in regression, where the second summand is the analogue of a regularization term for the weights and biases of e.g. a neural network.

Gradient-based optimizers can then be used to efficiently obtain $\hat{\theta}$ in Eq. (5), iteratively updating the parameter vector θ_m given the gradient of the loss function, denoted by $\nabla_\theta L_{\mathcal{M}}$, to navigate the surface defined by $L_{\mathcal{M}}$ with the aim to find the global minimum where $\nabla_\theta L_{\mathcal{M}}(\hat{\theta}) = 0$. As an example, the plain vanilla gradient descent algorithm is given by

$$\theta_{m+1} = \theta_m - \gamma \nabla_\theta L_{\mathcal{M}}(\theta_m) \tag{7}$$

where γ is the learning rate. Other gradient-based algorithms, such as the ADAM optimizer used in the section below, employ more advanced updating strategies to avoid convergence to local minima but stay in the spirit of Eq. (7).

2.1.3 Information indigestion

A naive minimization of $L_{\mathcal{M}}(\theta)$ with gradient-based methods is likely to fail, as its associated surface cannot be navigated properly. As illustrated in Fig. 1**A**, the loss surface associated with models characterized by complex dynamics consists of multiple local minima, which cause problems of convergence in efforts to reach the global minimum (depicted by the orange curve in Fig. 1**B**). Furthermore, in a neighbourhood of the global minimum, the gradient of the loss function is very large (a "ravine" with almost vertical walls), leading the optimizer to overshoot the true parameter values (depicted by the green curve in Fig. 1**B**). In Section S1 we show in a general setting that these problematic features arise from the dynamical properties of ecosystem models: when the dynamics are chaotic or exhibit a limit cycle (as is often the case for ecological dynamics [57, 21, 22, 23]), the dynamical trajectories exhibit high sensitivity to the model parameters and ICs. This means that a small modification of the parameters or ICs leads to large divergences over time. The prevalence of large deviations causes discontinuities, appearing as many sub-optimal local minima on the loss surface. Moreover, the true minimum can only be found in a narrow ravine that becomes narrower as the number of data points increases. Such surfaces are hardly navigable with gradient descent methods, but since the behaviour of $L_{\mathcal{M}}(\theta)$ critically

depends on the time horizon, we reformulate Eq. (6) in the following section by splitting the time series into mini-batches with a short time horizon.

2.2 ML framework for ecosystem models

2.2.1 Description of the mini-batch method

We propose a mini-batch method that splits the data into mini-batches with a short time horizon. Under perturbed parameters and ICs, chaotic or limit cycle dynamics only diverge after some characteristic simulation time; by splitting the time series into small mini-batches, discontinuities that cause the poor navigability of the landscape are therefore avoided. By averaging the associated losses during the training, the mini-batch method regularizes the loss function and makes it possible to combine the information contained in independent time series.

Consider the availability of S independent time series, where each time series $s \in \{1, \dots, S\}$ contains $K^{(s)}$ observations $\mathbf{y}^{(s)}_{1:K^{(s)}}$. To improve the ill-behaviour of Eq. (6), we split each time series into $M^{(s)}$ mini-batches, each of which defines a loss denoted by $L^{(s,m)}_{\mathcal{M}}$. Averaging the losses $L^{(s,m)}_{\mathcal{M}}$ leads to a reformulation of the loss function in Eq. (6), yielding

$$
\begin{aligned}
L^{\star}_{\mathcal{M}}(\theta) &= \tfrac{1}{S}\sum_{s=1}^{S} \tfrac{1}{M^{(s)}} \sum_{m=0}^{M^{(s)}-1} L^{(s,m)}_{\mathcal{M}}(\theta) \\
&= \tfrac{1}{S}\sum_{s=1}^{S} \tfrac{1}{M^{(s)}} \sum_{m=0}^{M^{(s)}-1} \Big[\tfrac{1}{K^{(s,m)}} \sum_{k=1}^{K^{(s,m)}} \| y^{(s)}_{k+mK^{(s)}/M^{(s)}} - h\left(\mathcal{M}(t_{k+mK^{(s)}}, p, x_0^{(s,m)})\right) \|_{\Sigma_y} \\
&\quad + \| y^{(s)}_{mK^{(s)}/M^{(s)}} - h\left(x_0^{(s,m)}\right) \|_{\Sigma_{x_0}} + \|p_b - p\|_{\Sigma_p} \Big]
\end{aligned}
\tag{8}
$$

where $\theta = (p, x_0^{(1,1)}, \dots x_0^{(1,M_1)}, x_0^{(2,1)} \dots, x_0^{(S,M_S)})$ consists of the augmented parameter vector containing $\sum_{s=1}^{S}\left[M^{(s)}-1\right]$ additional ICs to be estimated and denoted by $x_0^{(s,m)}$, $K^{(s,m)} = \begin{cases} K^{(s)}/M^{(s)} + 1 & : m < M \\ K^{(s)}/M^{(s)} & : m = M \end{cases}$ indicates the number of points to include in the mth batch, and p_b designates the prior knowledge of the parameter values. We refer to Fig. 1**C** for a graphical representation of Eq. (8). In contrast to Eq. (6), the ICs for each batch are constrained by an extra term with weight $\Sigma_{x_0}^{-1}$, which is needed in practice for better convergence. The ICs $x_0^{(s,m)}$ are nuisance parameters that augment the dimensionality of the parameter space as they are inferred for each batch. Nonetheless, the efficiency of gradient-based methods, together with the suggested numerical implementation detailed in the section below, largely circumvent the additional cost associated with the augmented dimensionality. $K^{(s,m)}$ expresses the overlap between each batch of data for $m < M^{(s)}$ and ensures that all

the available information is utilized to constrain the parameter vector p. By smoothing the ruggedness of the loss surface (orange curve in Fig. 1**A**), the proposed mini-batch method yields an improved navigability (orange curve in Fig. 1**B**). We show analytically in Section S1 that using $L^{\star}_{\mathcal{M}}$ yields a more navigable loss surface than if $L_{\mathcal{M}}(\theta)$ is used in Eq. (6).

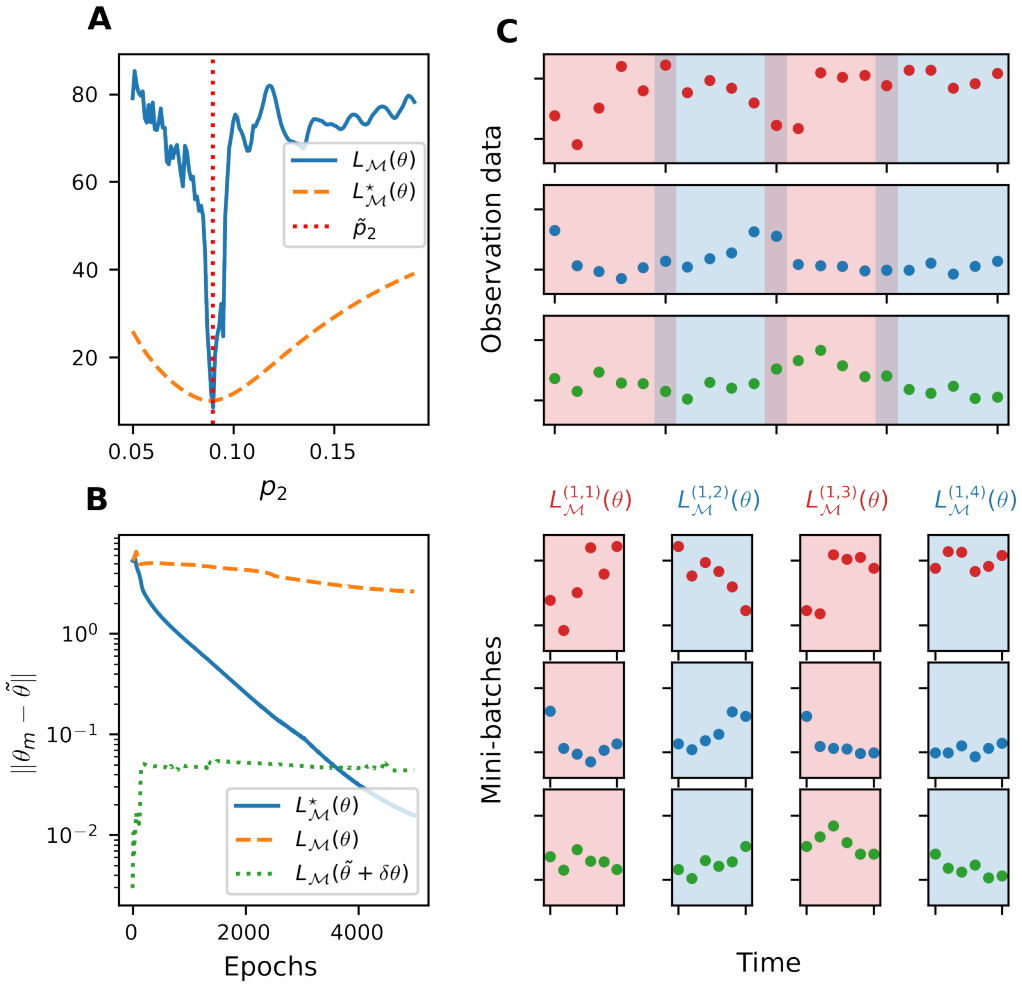

Figure 1: **Illustration of the mini-batch method.** **A** Characteristic features of the naive loss function $L_{\mathcal{M}}(\theta)$ and the mini-batch loss function $L^{\star}_{\mathcal{M}}(\theta)$. The blue and orange lines correspond to cross sections of $L_{\mathcal{M}}(\theta)$ and $L^{\star}_{\mathcal{M}}(\theta)$, respectively, obtained from the ecosystem model presented in the section Simulated food-web model as a case study. While the cross section of $L_{\mathcal{M}}(\theta)$ presents many local minima and a very large gradient in the neighbourhood of the true parameter, which renders the navigability of the loss surface difficult, the cross section of $L^{\star}_{\mathcal{M}}(\theta)$ is smooth and shows a single minimum, illustrating the regularization induced by the mini-batch method. **B** Convergence of gradient descent algorithms applied to $L_{\mathcal{M}}$ and $L^{\star}_{\mathcal{M}}$. The blue, orange and green lines correspond to the loss function evaluated against the epochs (number of parameter updates) using $L_{\mathcal{M}}$, $L_{\mathcal{M}}$ starting from a parameter value close to the true parameters, and $L^{\star}_{\mathcal{M}}$, respectively. The ill-behaviour of $L_{\mathcal{M}}$ leads to the convergence to a local minimum, while $L^{\star}_{\mathcal{M}}$ is associated with a smooth convergence to the true parameters. **C** Graphical representation of the proposed mini-batch method. To improve the navigability of the posterior landscape, the algorithm splits the time series into mini-batches with short time horizons (blue and red portions of the time series). Since mini-batches are treated independently, the method naturally extends to independent time series.

2.2.2 Numerical implementation of the ML framework

The choice of the optimizer and the correct calculation of the model sensitivity to the parameters and ICs, upon which the gradient of the loss function $\nabla L^{\star}_{\mathcal{M}}(\theta)$ depends, play an essential role in the success of the minimization of the loss function and the subsequent correct estimation of the MAP. To accelerate the learning process and make it more robust, we propose to combine the mini-batch method with modern optimizers and automatic differentiation. Building upon the software ecosystem SciML [53], we use **DifferentialEquations.jl** for the forward integration of the ecosystem model, as it provides highly efficient ODE solvers [58]. **DifferentialEquations.jl** is additionally compatible with automatic differentiation and includes an extensive set of sensitivity analysis methods [59], enabling the automatic generation of the model sensitivity to the parameters and ICs and guaranteeing their accuracy. This automatic generation greatly reduces the effort and potential errors associated with the adjoint code construction, enabling continuous development of the models. The accuracy is also an essential feature, as the model sensitivities are critically involved in the minimization of Eq. (6) and their inaccuracies can compromise the convergence of the gradient-based optimizers [60]. The interoperability of the SciML ecosystem further makes it possible to benefit from the tooling of the deep learning library **Flux.jl** and the nonlinear optimization library **Optim.jl** [61], providing state-of-the-art optimizers that are computationally efficient and well suited for highly parameterized models [62]. We use the adaptive, momentum-based Adam optimizer [63] to converge in the basin of attraction of the true parameters, which we substitute with the limited memory Broyden–Fletcher–Goldfarb–Shanno optimizer (L-BFGS) [64] for the final training epochs to ensure faster and more accurate convergence.

The reformulation of the learning problem in Eq. (8), together with the numerical implementation suggested above, define the proposed ML framework, which we benchmark with a concrete case scenario in the next section.

3 Simulated food-web model as a case study

We evaluate the performance of the ML framework by considering a food-web ecosystem composed of three functional compartments including a resource, consumers and predators. We use a reference model to generate the observation data and first assume a perfect-model setting, evaluating the performance of the ML framework in parametrizing the reference model for different noise levels, with incomplete observations, and with an increasing number of independent time series. Second, we relax the perfect-model assumption by considering two plausible candidate models capturing contrasting hypotheses regarding the ecological processes, and test

whether the ML framework can provide support for the true generating model by combining it with information-based model selection.

3.1 Three-compartment food-web ecosystem

We use a reference food-web model investigated in [65, 66, 67, 68] where a resource R is eaten by consumers C, which in turn are fed upon by predators P (model $\mathcal{M}_1$ in Fig. 2). We further consider an "omnivory variant" of the reference model introduced in [69], where predators are omnivorous and can feed upon the resource with a determined strength ω (model $\mathcal{M}_2$ in Fig. 2). These models generate fluctuations that resemble the behaviour of observed ecological time series [57], they produce chaotic dynamics that are notoriously challenging to forecast for a wide range of realistic parameters [70], and they have been used as benchmarks for proposed ecosystem forecasting methods (see [71, 12, 11]).

After nondimensionalization, the three-compartment model and the omnivory variant comprise a total of six and nine parameters, respectively: the mass-specific metabolic rate of consumers and predators x_C and x_P, the ingestion rate per unit metabolic rate of consumers and predators y_C and y_P (decomposed into y_{PC}, y_{PR} for the omnivory variant), the half saturation densities for the type II functional responses of the consumers and predators R_0 (decomposed into R_0, R_{02} for the omnivory variant) and C_0, and the omnivory strength ω for the omnivory variant (see Section S2 for the ODE details). Time is nondimensionalized by the resource growth rate and set to the biologically realistic value of 100% biomass increase per day, so that one unit of time corresponds to one day. The parameters in the simulations are set to the biologically realistic values proposed by [67, 69], which additionally ensure that the dynamics of the system are chaotic or show oscillations (see Section S2 for details).

We generate the observation data by sampling the simulated ecosystem dynamics and by contaminating the samples with noise. The noise variance–covariance matrix Σ_y is set to be diagonal, with entries that are proportional to the sample variances of the observables

$$\operatorname{diag} \Sigma_y = r^2[\operatorname{Var}(\tilde{y}_1), \ldots, \operatorname{Var}(\tilde{y}_d)] \tag{9}$$

where r indicates the noise level and $\tilde{y}$ corresponds to the noiseless data generated with the true parameter values $\tilde{\theta}$. We sample the simulated dynamics after a long burn-in time ($t > 500$) to ensure that transient dynamics are not observed. A visual representation of the generated data is displayed in Fig. 2**B**. We assume a uniform distribution of the parameter priors, and randomly draw initial parameter estimates from a uniform

distribution so that the initial parameter estimates follow $\mathcal{U}(0, 2\tilde{p})$.

We consider two different settings: one where all compartment abundances are observable, i.e. the observing system map h is the identity, and one where only predator and consumer abundances are available, i.e. discarding the resource abundance data. For both settings, structural identifiability is tested with the Julia library **StructuralIdentifibility.jl** [72] and verified globally. This means that in theory, the unique observation of predator and consumer abundances carries the information required for a complete characterization of all the model parameters.

For the meta-parameters of the ML framework, we set $\Sigma_{x_0} = \frac{M^{(1)}}{K^{(1)}}\Sigma_y$, we use Adam with $\gamma_m = \mathbf{1}_{[0,2000]}(m)10^{-1} + \mathbf{1}_{[0,2000]}(m)10^{-2} + \mathbf{1}_{[0,2000]}(m)10^{-3}$, $\beta_1 = 0.9$ and $\beta_2 = 0.999$ for the first 6000 epochs, where γ_m corresponds to the learning rate of the mth epoch, and we use L-BFGS for the last 200 epochs.

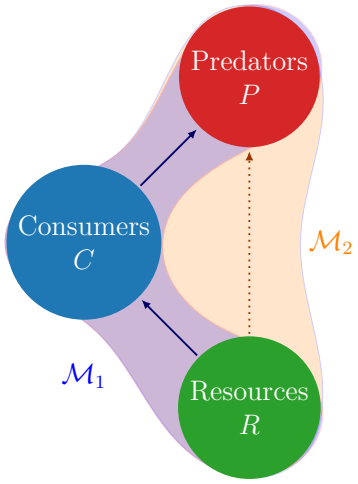

Figure 2: **Reference food web systems considered.** In Section 3.2, the blue model $\mathcal{M}_1$ from [65] is considered, where a resource is eaten by consumers, which are themselves eaten by predators. In Section 3.3, the orange model $\mathcal{M}_2$, corresponding to the omnivory variant introduced in [69], is also considered.

3.2 Parameter learning in a perfect-model setting

We generate time series from the reference food-web model (blue model in Fig. 2) under varying noise levels and for both settings with total and partial observations, sampling from the model simulations every four days (see Fig. 1**C** for an illustration of the generated data). We then apply the ML method to the generated data, with a focus on evaluating its performance in recovering the true parameters $\tilde{p}$ and on its forecast skill.

We evaluate the performance in recovering the true parameters with two metrics, namely the coefficient of determination between the true parameters $\tilde{x}_P$ and the estimated parameters $\hat{x}_P$, denoted by $R^2_{x_P}$, and the

relative parameter error for the ensemble of training simulations, denoted by $|(\hat{p}-\tilde{p})/\tilde{p}|$ and calculated as the median relative parameter error across the six estimated parameters. To evaluate the out-of-sample forecast skill, we simulate the model beyond the training time span by using the estimated ICs of the last batch of each independent time series. Further, we quantify the forecast skill, denoted by ρ^2, by computing the mean squared correlation across all the independent time series between the prey abundance generated with the true parameter $\tilde{\theta}$ and the predicted prey abundance. We obtain summary statistics of the metrics by varying the critical parameter value x_P, generating a total of 50 simulations for each noise level and setting considered. While only x_P is varied, all the parameters together with the ICs are collectively fitted.

3.2.1 Robustness of the ML framework against noise and incomplete observations

We set the number of time series to $S = 1$, the time series length to $K = 80$, and the number of mini-batches to $M = 8$. We investigate the ML framework performance against observational noise and under the setting with complete or partial observations.

In the complete observation setting, the ML framework can very accurately recover the true parameter values under a moderate observational noise, with mean $|(\hat{p}-\tilde{p})/\tilde{p}| = 6\%$ and $R^2_{x_P} = 0.99$ for 20% observational noise ($r = 0.2$; see Fig. 3**A**, red dots). In the partial observation setting, fair results are also obtained with $|(\hat{p}-\tilde{p})/\tilde{p}| = 16\%$ and $R^2_{x_P} = 0.89$ (Fig. 3**A**, blue triangles). On top of being accurate, the ML framework further shows a very short inference time, i.e. 36 seconds in the complete observation setting and 34 seconds in the partial observation setting (see Table S1 for details). To investigate systematically how the ML framework accommodates different levels of observational noise, we vary the noise level from $r = 0.$ to $r = 1$ and calculate $|(\hat{p}-\tilde{p})/\tilde{p}|$. Results reported in Fig. 3**B** show that the ML framework handles observational noise well and the response of the logarithm of $|(\hat{p}-\tilde{p})/\tilde{p}|$ to r only differs by a constant factor between the complete and partial observation settings.

Given that the food-web dynamics are chaotic (Section S2), excellent performance in parameter estimation might not be sufficient to provide accurate forecasts. We therefore also test the ML algorithm by evaluating how the forecast skill ρ^2 is affected by the noise level under complete or partial observations. Results for the complete and partial observation settings reported in Fig. 3**C** show good forecast skill under moderate observational noise, where ρ^2 linearly decreases with the time horizon considered.

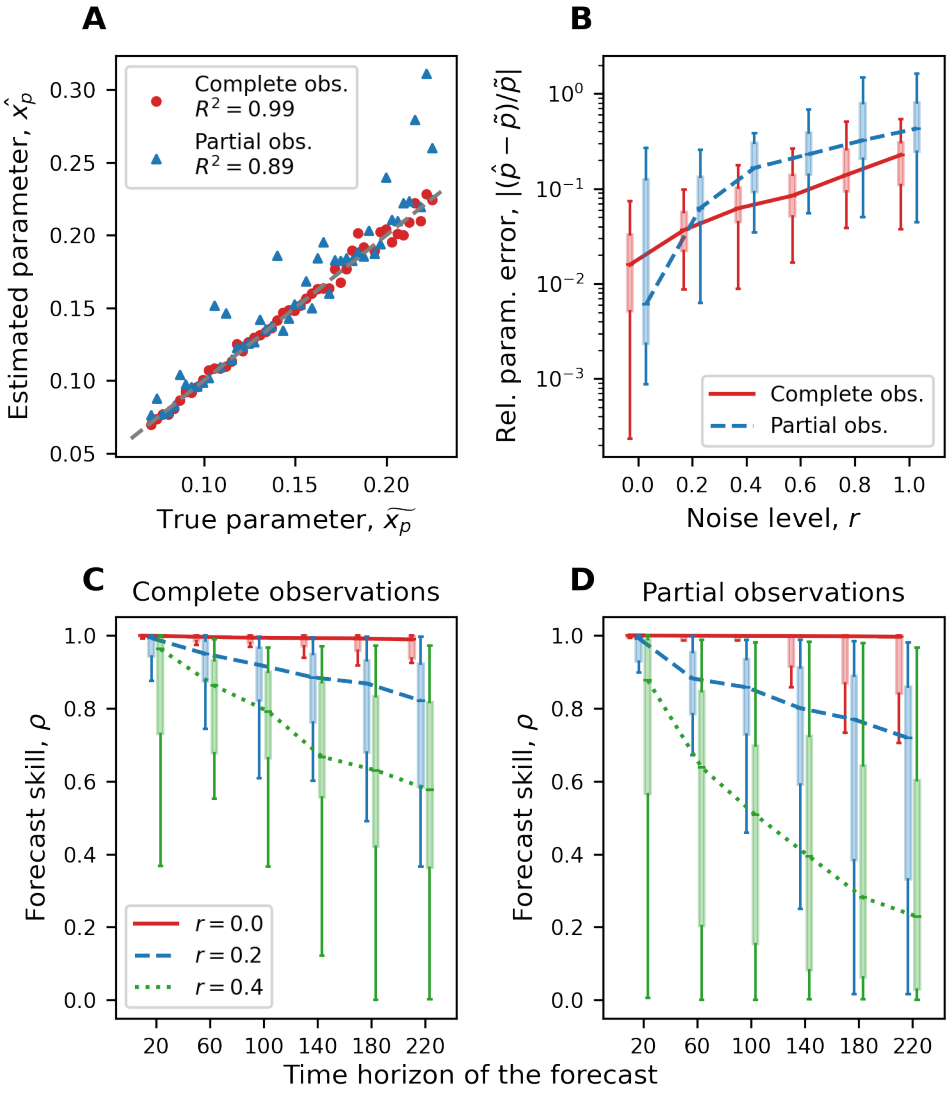

Figure 3: **Performance of the proposed ML framework for varying noise levels, under the complete and the partial observation setting.** **A** True parameter $\widetilde{x_P}$ against estimated parameter $\hat{x_P}$ for $r = 0.2$ under the complete and the partial observation setting. Although the parameter estimation is more accurate when complete observations of abundance are available, parameters show a fair fit when estimated with partial observations. **B** Relative parameter error $|(\hat{p}-\tilde{p})/\tilde{p}|$ for varying noise levels under the setting of complete or partial observations. **B** supports the above observation for varying noise levels. **C** Forecast skill ρ^2 of the trained model under the complete observation setting. **D** Analogous data under the partial observation setting. In **A**–**D**, the batch size is set to $m = 6$.

3.2.2 Capability of the ML framework to harness multiple time series

We further investigate the ability of the ML framework to process and combine information from independent datasets. We reduce the time horizon of the observation data by setting the time series length to $K = 12$, generate two datasets comprising $S = 1$ and $S = 6$ independent time series, respectively, – obtained from independent ICs – and set the number of batches for each time series s to $M^{(s)} = 2$. In both the complete and partial observation settings, we find that the relative parameter error $|(\hat{p}-\tilde{p})/\tilde{p}|$ is consistently lower in the simulations with a larger number of time series (Fig. 4**A**). The forecast skill is also consistently improved as more independent time series are processed (Fig. 4**B**), and the forecast skill for long-term predictions considerably increases. These results confirm the robustness of the ML framework against noise and partial observations, and show that the ML framework can efficiently harness the information from disparate observation datasets.

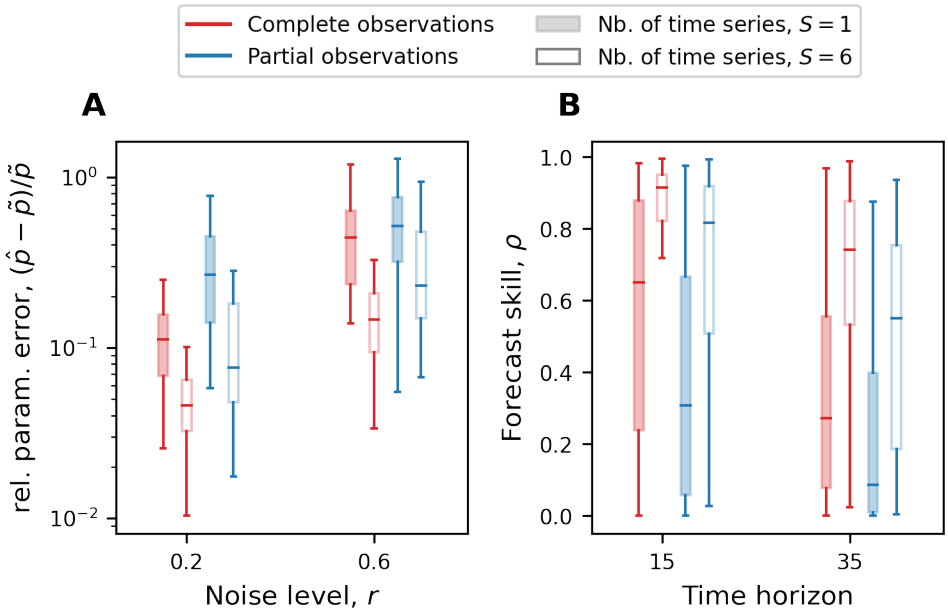

Figure 4: **Performance of the ML framework in processing and combining the information of multiple independent data sets. A** Relative parameter error $|(\hat{p}-\tilde{p})/\tilde{p}|$ for different numbers of time series and levels of noise, under the complete and the partial observation setting. **B** Forecast skill for different numbers of time series and time horizons of the forecasts, under the complete and the partial observation setting. $r = 0.1$. In **A**–**B**, $|(\hat{p}-\tilde{p})/\tilde{p}|$ decreases while ρ^2 increases as the number of time series processed increases, demonstrating the capacity of the ML framework to process and combine the information from independent time series. In **A**–**B**, each box plot corresponds to 100 independent simulations where $\tilde{x_P}$ varies $\tilde{x_P} \in [0.071, 0.225]$, and the batch size is set to $m = 6$.

3.3 Elucidating mechanistic pathways

Finally, we relax the perfect-model assumption and investigate whether the ML framework can provide statistical support for the true generating model among several candidates with information-based model selection. Specifically, we investigate whether the ML framework can detect omnivory from single observations of time series. We generate multiple observation datasets from the omnivory variant model $\mathcal{M}_2$ for different omnivory strengths ω and noise levels r. We consider both the standard model $\mathcal{M}_1$ and the omnivory variant model $\mathcal{M}_2$ as two plausible candidate models (see Fig. 2 for a graphical illustration of the models). We use the Akaike information criterion (AIC) to select the model with the strongest support in relation to the data [73]. In the specific case of our framework, we calculate the AIC as

$$\mathrm{AIC}_{\mathcal{M}_i} = -2\ln(p(\hat{\theta}, M_i|\mathbf{y}_{1:K})) + 2k_{\mathcal{M}_i} \tag{10}$$

where $p(\hat{\theta}, \mathcal{M}_i|y_{1:K})$ corresponds to the maximum value of the likelihood of the model $\mathcal{M}_i$ given the data, and $k_{\mathcal{M}_i}$ is the number of parameters in the model $\mathcal{M}_i$. The AIC ranks the most probable models by penalizing complexity to balance information loss and parsimony, where candidate models with the lowest scores are ranked as the most likely. We consider the Akaike weights $w_{\mathcal{M}_i} = \frac{\exp(-\Delta\mathrm{AIC}_{\mathcal{M}_i}/2)}{\sum_j \exp(-\Delta\mathrm{AIC}_{\mathcal{M}_j}/2)}$, where $\Delta\mathrm{AIC}_{\mathcal{M}_i} = \mathrm{AIC}_{\mathcal{M}_i} - \min_j \mathrm{AIC}_{\mathcal{M}_j}$, which can be directly interpreted as the probability that $\mathcal{M}_i$ is the most appropriate model given the data (see [74]). We expect that the Akaike weights provide support for the generating model $\mathcal{M}_2$ only across values where $\omega > 0$, as $\mathcal{M}_1$ is equivalent to $\mathcal{M}_2$ when $\omega = 0$ and $\mathcal{M}_2$ is penalized by its three additional parameters.

In the complete observation setting, for moderate observational noise ($r = 0.1$) we find that $\mathcal{M}_1$ is given strong support for $\omega < 0.07$ ($w_{\mathcal{M}_1} > 98\%$) and that $\mathcal{M}_2$ if favored for $\omega > 0.08$ ($w_{\mathcal{M}_2} > 99\%$), providing overall strong support for the true model over a large range of ω values (Fig. 5**A**). As the observational noise increases the support strength naturally decreases, leading to an increased range of ω values where the simplest model $\mathcal{M}_1$ is favored or where no model is given strong support (Fig. 5**B**-**C**). On the other hand, in the partial observation setting, the lack of data prevents the correct estimation of the omnivory variant model parameters, leading model $\mathcal{M}_1$ to be supported for an even larger range of ω values (Fig. S1).

Overall, the ML framework provides statistical support for the model embedding the most appropriate hypotheses given the available data. With appropriate data, the proposed ML framework can therefore elucidate mechanistic pathways and infer ecological processes by utilizing information-based model selection.

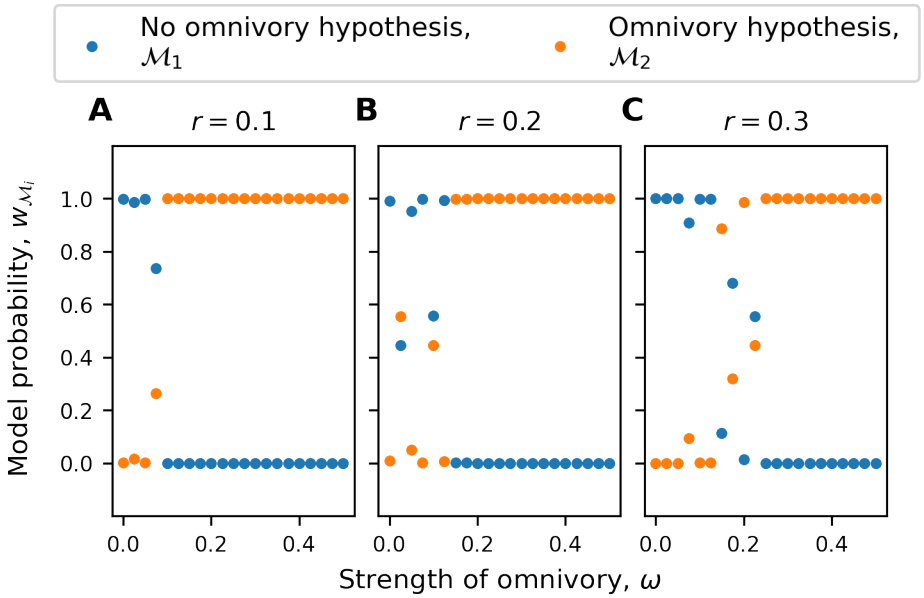

Figure 5: **Performance of the ML framework in supporting the predator omnivory hypothesis in a food web. A–C** Hypothesis testing for levels of noise $r = 0.1, 0.2, 0.3$. Blue dots correspond to $w_{\mathcal{M}_1}$, the Akaike weights of the simple food-web model, and orange dots correspond to $w_{\mathcal{M}_2} = 1 - w_{\mathcal{M}_1}$, the Aikaike weights of the omnivory model, which can be interpreted as model probabilities. **A–C** indicate that the ML framework can detect omnivory, as the omnivory model $\mathcal{M}_2$ is given strong support ($w_{\mathcal{M}_2} > 99\%$) for most of the ω range investigated.

4 Discussion

We propose a ML framework combining a mini-batch method inspired by multiple shooting methods [52] with automatic differentiation [53] and state-of-the-art variational optimizers [63] to efficiently and accurately parametrize complex dynamical models. We show formally that splitting the data into mini-batches with a short time horizon regularizes the loss function associated with dynamical models characterized by complex dynamics, such as chaotic dynamics and limit cycles (Section S1). We demonstrate numerically that this reformulation ensures the success of gradient-based optimizers to parametrize ecosystem models (Figs. 1, 3 and 4). This mini-batch method is also relevant beyond variational methods and applies to any inferential method navigating the posterior landscape, such as evolutionary algorithms [75, 76] or Markov Chain Monte Carlo methods [27, 28, 29, 30, 31]. The proposed approach is particularly relevant for the parametrization of ecosystem models incorporating realistic ecological and adaptive mechanisms [13], which are generally associated with strong nonlinearities due to the complexity of processes linking interacting ecological compartments [57, 21, 22, 23]. It further integrates the practical constraints of available ecological datasets [77], accommodating incomplete, noisy, shallow and independent observation data. Overall, the ML framework successfully blends ML methods

with mechanistic ecosystem models to learn from ecological time series, and it could therefore improve our quantitative understanding of ecosystem dynamics and help to anticipate their responses to global changes [13].

Our work contributes to the ongoing effort to better assimilate observational data into mechanistic models [20, 78, 49], with a specific focus on the parametrization of ecosystem models with strong nonlinearities. Recently, [79] proposed an alternative framework dubbed "systems biology informed deep learning", where a neural network is fitted to the data and the additional mechanistic model constraints are integrated. This alternative framework extends previous colocation methods [80, 81] and has the advantage of being able to parametrize stochastic models. As it requires the selection of a neural network architecture and a "goodness of fit" parameter, it nevertheless imposes an additional layer of complexity, which might negatively affect the model parametrization [79]. In contrast, the ML framework proposed here trains the model directly against data, using automatic differentiation and sensitivity analysis in order to apply variational optimizers directly to the model simulations. This makes it possible to bypass the use of neural networks, rendering the parametrization process simpler and more amenable to model selection [80].

By integrating the practical constraints imposed by ecological datasets, the ML framework can learn from short time series with partial and noisy observations (Figs. 3 and 4). Local ecosystem surveys, such as marine trawling surveys or local terrestrial surveys ([82, 77, 83] and references therein), provide time series that are generally shallow in time but composed of many replicates [84, 85], in part due to the practical difficulties of long-term monitoring [11]. Our results show that the inclusion of multiple independent time series in the training dataset reduces the error in the parameter estimates and increases the forecast skill (Fig. 4). This indicates that the proposed ML framework could, in practice, efficiently harness the information available in current ecological datasets. Instead of directly comparing simulated and observed data, matching time-averaged statistics between observations and simulations (e.g. means and covariances) could further yield an improved assimilation of observations from diverse data sources, such as global observations of productivity from satellites and local surveys, as proposed for climate models [55]. Overall, the proposed ML framework accommodates the specificities of current ecological time series and can improve the assimilation of ecological data into mechanistic ecosystem models.

Our work can help elucidate mechanistic pathways by contrasting hypotheses embedded in model variants. Using information-criterion-based model selection, we demonstrate with a case study that the ML framework is able to provide statistical support for the true generating model among two different candidates (Fig. 5). Importantly, the ML framework can perform model selection on complex models, incorporating key mechanisms such as trait–species interactions, evolutionary potential and responses to environmental conditions, which have

been shown to be important in mediating ecosystem dynamics and must be refined in models to improve predictive accuracy [13]. The ML framework can therefore lead to the improvement of current ecosystem models and knowledge, which is crucially needed given that key ecological processes are only partially described in most ecosystem models [20]. AIC can also be used to ensure the interpretability of the model parameters, and should be preferred to estimating the parameter uncertainty through e.g. the Cramer Rao inqequality [74]: by favouring models with less complexity, model selection techniques disqualify uninformative parameters to ensure interpretability [74]. This has the extra benefit of reducing the dimensionality of the parameter space, hence improving the estimation of other parameters. Following recent novel approaches to investigate ecological hypotheses [45], our method contributes to the development of a process understanding of ecosystem functions and provides a path forward to better link ecological theory and data.

The proposed approach still presents a set of limitations, which might hamper its success under specific situations. First, while the use of mini-batches smooths the loss surface and ensures better convergence, it also flattens the loss surface around the true parameter value, which consequently deteriorates the precision of the inferred parameters because the loss function takes similar values in an extended neighbourhood of the true parameters. To circumvent this issue, iterative training can be performed, where the learning is initiated by a short batch length $K^{(s)}$ to identify the region with the most probable parameters, and in subsequent iterations the batch length is increased to improve the precision of the inference. Iterative training could also improve the lack of statistical support obtained in hypothesis testing experiments (see simulations in Figs. S1 and 5 where none of the models is given statistical support), as it would increase differences in likelihood for parameter values around the neighbourhood of the true parameter values. Second, our results highlight that the data might not provide enough constraints for a correct parametrization (Fig. 4, partial observation setting and $S = 1$). Pre-experimental analyses with simulated synthetic data might therefore be required to design the sampling protocol and campaign to ensure an adequate sampling effort [86, 87]. Third, while in Eq. (8) it is assumed that the parameter values are the same across the time series, strong regional variability might also be observed among the spatially replicated data, causing parameter values to vary across the replicates. The knowledge of this variability could motivate partial pooling [88] or parametrization of the parameters in terms of environmental conditions [89], to account for the independence of the parameter values across the replicates. Finally, while the proposed ML framework greatly improves convergence, it could still be that – even with a large amount of data – poor initial parameter estimates, a large number of free parameters, or high noise levels prevent convergence to the true minimum. Performing multiple runs with varying initial parameter estimates can ensure that the maximum a priori estimate is reliable. If this is not the case, stochasticity could further

be introduced within the ML framework to prevent the convergence to local minima, where only a subset of mini-batches are fitted at each epoch [90].

5 Conclusion

We proposed a ML framework based on a mini-batch method combined with automatic differentiation and state-of-the-art optimizers to estimate the parameters and improve the forecast skill of complex ecosystem models from observation data. The ML framework was benchmarked with a realistic ecosystem model characterized by strong nonlinearities and delivered excellent performance, accommodating the practical constraints imposed by the quality and availability of ecological datasets. Our experiments have further illustrated the ability of the ML framework to discriminate between several candidate models, enabling the testing of ecological theories against data and the improvement of current mechanistic models. Given the increasing number of ecological datasets following the development of monitoring technologies such as environmental DNA [91], remote sensing [92], bioaccoustics [93], and citizen observations [94], the proposed ML framework opens up new opportunities for the quantitative investigation of current ecosystem functions [45] and the prediction of ecosystem responses to increasing disruptions [13].

6 Acknowledgements

L.P. and V.B. were supported by the SNF grant 310030E_205556. P.V.A. was supported by an ETH postdoctoral fellowship.

7 Competing interests

The authors declare no competing interests.

8 Code availability

The ML framework is implemented in the multi-purpose Julia package **MiniBatchInference.jl** available at https://github.com/vboussange/MiniBatchInference.jl, and the simulation code is available at https://github.com/vboussange/mini-batching-ecological-data.

9 Contributions

Conceptualization V.B. and P.V.

Formal Analysis V.B. and P.V.

Funding Acquisition P.V. and L.P.

Investigation V.B. and P.V.

Methodology V.B.

Software V.B.

Validation V.B.

Visualization V.B.

Writing – Original Draft Preparation V.B., P.V. and L.P.

References

[1] Gordon B. Bonan. Forests and climate change: forcings, feedbacks, and the climate benefits of forests. *Science*, 320(5882):1444–1449, jun 2008.

[2] Claire Kremen. Managing ecosystem services: what do we need to know about their ecology? *Ecology Letters*, 8(5):468–479, may 2005.

[3] Scott C. Doney. The growing human footprint on coastal and open-ocean biogeochemistry. *Science*, 328(5985):1512–1516, jun 2010.

[4] Erle C. Ellis. Anthropogenic transformation of the terrestrial biosphere. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 369(1938):1010–1035, mar 2011.

[5] Guy Midgley and Lee Hannah. Extinction Risk from Climate Change. In *Biodiversity and Climate Change*, pages 294–296. Yale University Press, 2019.

[6] Ian L. Boyd. The art of ecological modeling. *Science*, 337(6092):306–307, jul 2012.

[7] Marten Scheffer, Steve Carpenter, Jonathan A Foley, Carl Folke, and Brian Walker. Catastrophic shifts in ecosystems. *Nature*, 413(6856):591–596, oct 2001.

[8] Anthony D. Barnosky, Elizabeth A. Hadly, Jordi Bascompte, Eric L. Berlow, James H. Brown, Mikael Fortelius, Wayne M. Getz, John Harte, Alan Hastings, Pablo A. Marquet, Neo D. Martinez, Arne Mooers, Peter Roopnarine, Geerat Vermeij, John W. Williams, Rosemary Gillespie, Justin Kitzes, Charles Marshall, Nicholas Matzke, David P. Mindell, Eloy Revilla, and Adam B. Smith. Approaching a state shift in Earth's biosphere. *Nature*, 486(7401):52–58, 2012.

[9] Benjamin Deneu, Maximilien Servajean, Pierre Bonnet, Christophe Botella, François Munoz, and Alexis Joly. Convolutional neural networks improve species distribution modelling by capturing the spatial structure of the environment. *PLOS Computational Biology*, 17(4):e1008856, apr 2021.

[10] Hao Ye, Richard J. Beamish, Sarah M. Glaser, Sue C. H. Grant, Chih-hao Hsieh, Laura J. Richards, Jon T. Schnute, and George Sugihara. Equation-free mechanistic ecosystem forecasting using empirical dynamic modeling. *Proceedings of the National Academy of Sciences of the United States of America*, 112(13):E1569–E1576, mar 2015.

[11] Hao Ye and George Sugihara. Information leverage in interconnected ecosystems: Overcoming the curse of dimensionality. *Science*, 353(6302):922–925, aug 2016.

[12] Ethan R. Deyle, Robert M. May, Stephan B. Munch, and George Sugihara. Tracking and forecasting ecosystem interactions in real time. *Proceedings of the Royal Society B: Biological Sciences*, 283(1822):20152258, jan 2016.

[13] M. C. Urban, G. Bocedi, A. P. Hendry, J.-B. Mihoub, G. Pe'er, A. Singer, J. R. Bridle, L. G. Crozier, L. De Meester, W. Godsoe, A. Gonzalez, J. J. Hellmann, R. D. Holt, A. Huth, K. Johst, C. B. Krug, P. W. Leadley, S. C. F. Palmer, J. H. Pantel, A. Schmitz, P. A. Zollner, and J. M. J. Travis. Improving the forecast for biodiversity under climate change. *Science*, 353(6304), sep 2016.

[14] William L. Geary, Michael Bode, Tim S. Doherty, Elizabeth A. Fulton, Dale G. Nimmo, Ayesha I. T. Tulloch, Vivitskaia J. D. Tulloch, and Euan G. Ritchie. A guide to ecosystem models and their environmental applications. *Nature Ecology & Evolution*, 4(11):1459–1471, nov 2020.

[15] Jon Norberg, Mark C. Urban, Mark Vellend, Christopher A. Klausmeier, and Nicolas Loeuille. Eco-evolutionary responses of biodiversity to climate change. *Nature Climate Change*, 2(10):747–751, oct 2012.

[16] Donald L. DeAngelis and Simeon Yurek. Equation-free modeling unravels the behavior of complex ecological systems. *Proceedings of the National Academy of Sciences*, 112(13):3856–3857, mar 2015.

[17] Rosie A. Fisher, Charles D. Koven, William R.L. Anderegg, Bradley O. Christoffersen, Michael C. Dietze, Caroline E. Farrior, Jennifer A. Holm, George C. Hurtt, Ryan G. Knox, Peter J. Lawrence, Jeremy W. Lichstein, Marcos Longo, Ashley M. Matheny, David Medvigy, Helene C. Muller-Landau, Thomas L. Powell, Shawn P. Serbin, Hisashi Sato, Jacquelyn K. Shuman, Benjamin Smith, Anna T. Trugman, Toni Viskari, Hans Verbeeck, Ensheng Weng, Chonggang Xu, Xiangtao Xu, Tao Zhang, and Paul R. Moorcroft. Vegetation demographics in Earth System Models: A review of progress and priorities. *Global Change Biology*, 24(1):35–54, 2018.

[18] Ryan N. Gutenkunst, Joshua J. Waterfall, Fergal P. Casey, Kevin S. Brown, Christopher R. Myers, and James P. Sethna. Universally sloppy parameter sensitivities in systems biology models. *PLOS Computational Biology*, 3(10):e189, oct 2007.

[19] Simon Scheiter, Liam Langan, and Steven I. Higgins. Next-generation dynamic global vegetation models: learning from community ecology. *New Phytologist*, 198(3):957–969, may 2013.

[20] Markus Schartau, Philip Wallhead, John Hemmings, Ulrike Löptien, Iris Kriest, Shubham Krishna, Ben A. Ward, Thomas Slawig, and Andreas Oschlies. Reviews and syntheses: parameter identification in marine planktonic ecosystem modelling. *Biogeosciences*, 14(6):1647–1701, mar 2017.

[21] Alan Hastings, Carole L. Hom, Stephen Ellner, Peter Turchin, and H. Charles J. Godfray. Chaos in Ecology: Is mother nature a strange attractor? *Annual Review of Ecology and Systematics*, 24(1):1–33, nov 1993.

[22] Jef Huisman and Franz J. Weissing. Biodiversity of plankton by species oscillations and chaos. *Nature*, 402(6760):407–410, nov 1999.

[23] Elisa Benincà, Jef Huisman, Reinhard Heerkloss, Klaus D. Jöhnk, Pedro Branco, Egbert H. Van Nes, Marten Scheffer, and Stephen P. Ellner. Chaos in a long-term experiment with a plankton community. *Nature*, 451(7180):822–825, feb 2008.

[24] M. Gehlen, R. Barciela, L. Bertino, P. Brasseur, M. Butenschön, F. Chai, A. Crise, Y. Drillet, D. Ford, D. Lavoie, P. Lehodey, C. Perruche, A. Samuelsen, and E. Simon. Building the capacity for forecasting marine biogeochemistry and ecosystems: recent advances and future developments. *Journal of Operational Oceanography*, 8(sup1):s168–s187, apr 2015.

[25] Drew Purves, Jörn P. W. Scharlemann, Mike Harfoot, Tim Newbold, Derek P. Tittensor, Jon Hutton, and Stephen Emmott. Time to model all life on Earth. *Nature*, 493(7432):295–297, jan 2013.

[26] Wendy Gentleman, Andrew Leising, Bruce Frost, Suzanne Strom, and James Murray. Functional responses for zooplankton feeding on multiple resources: a review of assumptions and biological dynamics. *Deep Sea Research Part II: Topical Studies in Oceanography*, 50(22-26):2847–2875, nov 2003.

[27] Risto Lignell, Heikki Haario, Marko Laine, and T. Frede Thingstad. Getting the "right" parameter values for models of the pelagic microbial food web. *Limnology and Oceanography*, 58(1):301–313, jan 2013.

[28] Steven I. Higgins, Simon Scheiter, and Mahesh Sankaran. The stability of African savannas: insights from the indirect estimation of the parameters of a dynamic model. *Ecology*, 91(6):1682–1692, jun 2010.

[29] Tao Xu, Luther White, Dafeng Hui, and Yiqi Luo. Probabilistic inversion of a terrestrial ecosystem model: Analysis of uncertainty in parameter estimation and model prediction. *Global Biogeochemical Cycles*, 20(2):n/a–n/a, jun 2006.

[30] J. Fiechter, R. Herbei, W. Leeds, J. Brown, R. Milliff, C. Wikle, A. Moore, and T. Powell. A Bayesian parameter estimation method applied to a marine ecosystem model for the coastal Gulf of Alaska. *Ecological Modelling*, 258:122–133, jun 2013.

[31] Benjamin Rosenbaum, Michael Raatz, Guntram Weithoff, Gregor F. Fussmann, and Ursula Gaedke. Estimating parameters from multiple time series of population dynamics using Bayesian inference. *Frontiers in Ecology and Evolution*, 6(JAN), jan 2019.

[32] Sanmitra Gosh, Paul Birrell, and Daniela De Angelis. Variational inference for nonlinear ordinary differential equations. *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, 130(29):2719–2727, 2021.

[33] Richard J. Matear. Parameter optimization and analysis of ecosystem models using simulated annealing: A case study at Station P. *Journal of Marine Research*, 53(4):571–607, jul 1995.

[34] Ben A. Ward, Marjorie A.M. Friedrichs, Thomas R. Anderson, and Andreas Oschlies. Parameter optimisation techniques and the problem of underdetermination in marine biogeochemical models. *Journal of Marine Systems*, 81(1-2):34–43, apr 2010.

[35] Laurent Bertino, Geir Evensen, and Hans Wackernagel. Sequential data assimilation techniques in oceanography. *International Statistical Review*, 71(2):223–241, aug 2003.

[36] Maéva Doron, Pierre Brasseur, Jean Michel Brankart, Svetlana N. Losa, and Angélique Melet. Stochastic estimation of biogeochemical parameters from Globcolour ocean colour satellite data in a North Atlantic 3D ocean coupled physical-biogeochemical model. *Journal of Marine Systems*, 117-118:81–95, 2013.

[37] M.E. Gharamti, J. Tjiputra, I. Bethke, A. Samuelsen, I. Skjelvan, M. Bentsen, and L. Bertino. Ensemble data assimilation for ocean biogeochemical state and parameter estimation at different sites. *Ocean Modelling*, 112:65–89, apr 2017.

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In I Guyon, U Von Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[39] Katja Fennel, Martin Losch, Jens Schröter, and Manfred Wenzel. Testing a marine ecosystem model: sensitivity analysis and parameter optimization. *Journal of Marine Systems*, 28(1-2):45–63, feb 2001.

[40] Y.H. Spitz, J.R. Moisan, M.R. Abbott, and J.G. Richman. Data assimilation and a pelagic ecosystem model: parameterization using time series observations. *Journal of Marine Systems*, 16(1-2):51–68, sep 1998.

[41] Yongjin Xiao and Marjorie A. M. Friedrichs. The assimilation of satellite-derived data into a one-dimensional lower trophic level marine ecosystem model. *Journal of Geophysical Research: Oceans*, 119(4):2691–2712, apr 2014.

[42] Joanna S. Pelc, Ehouarn Simon, Laurent Bertino, Ghada El Serafy, and Arnold W. Heemink. Application of model reduced 4D-Var to a 1D ecosystem model. *Ocean Modelling*, 57-58:43–58, nov 2012.

[43] Qing Zhu and Qianlai Zhuang. Ecosystem biogeochemistry model parameterization: Do more flux data result in a better model in predicting carbon flux? *Ecosphere*, 6(12):art283, dec 2015.

[44] John P. DeLong, Torrance C. Hanley, and David A. Vasseur. Predator-prey dynamics and the plasticity of predator body size. *Functional Ecology*, 28(2):487–493, apr 2014.

[45] Alva Curtsdotter, H. Thomas Banks, John E. Banks, Mattias Jonsson, Tomas Jonsson, Amanda N. Laubmeier, Michael Traugott, and Riccardo Bommarco. Ecosystem function in predator–prey food webs—confronting dynamic models with empirical data. *Journal of Animal Ecology*, 88(2):196–210, feb 2019.

[46] Attila Gábor and Julio R. Banga. Robust and efficient parameter estimation in dynamic models of biological systems. *BMC Systems Biology*, 9(1):74, dec 2015.

[47] Linda M. Lawson, Yvette H. Spitz, Eileen E. Hofmann, and Robert Bryan Long. A data assimilation technique applied to a predator-prey model. *Bulletin of Mathematical Biology*, 57(4):593–617, jul 1995.

[48] Jared Willard, Xiaowei Jia, Shaoming Xu, Michael Steinbach, and Vipin Kumar. Integrating scientific knowledge with machine learning for engineering and environmental systems. *ACM Computing Surveys*, 1(1):1–35, 2020.

[49] K. Kashinath, M. Mustafa, A. Albert, J. L. Wu, C. Jiang, S. Esmaeilzadeh, K. Azizzadenesheli, R. Wang, A. Chattopadhyay, A. Singh, A. Manepalli, D. Chirila, R. Yu, R. Walters, B. White, H. Xiao, H. A. Tchelepi, P. Marcus, A. Anandkumar, P. Hassanzadeh, and Prabhat. Physics-informed machine learning: Case studies for weather and climate modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379(2194), 2021.

[50] Mark Alber, Adrian Buganza Tepole, William R. Cannon, Suvranu De, Salvador Dura-Bernal, Krishna Garikipati, George Karniadakis, William W. Lytton, Paris Perdikaris, Linda Petzold, and Ellen Kuhl. Integrating machine learning and multiscale modeling—perspectives, challenges, and opportunities in the biological, biomedical, and behavioral sciences. *npj Digital Medicine*, 2(1):115, dec 2019.

[51] Grace C. Y. Peng, Mark Alber, Adrian Buganza Tepole, William R. Cannon, Suvranu De, Savador Dura-Bernal, Krishna Garikipati, George Karniadakis, William W. Lytton, Paris Perdikaris, Linda Petzold, and Ellen Kuhl. Multiscale modeling meets machine learning: What can we learn? *Archives of Computational Methods in Engineering*, 28(3):1017–1037, may 2021.

[52] V. F. Pisarenko and D. Sornette. Statistical methods of parameter estimation for deterministically chaotic time series. *Physical Review E*, 69(3):036122, mar 2004.

[53] Christopher Rackauckas, Yingbo Ma, Julius Martensen, Collin Warner, Kirill Zubov, Rohit Supekar, Dominic Skinner, Ali Ramadhan, and Alan Edelman. Universal Differential Equations for Scientific Machine Learning. jan 2020.

[54] Marc Bocquet, Julien Brajard, Alberto Carrassi, and Laurent Bertino. Data assimilation as a learning tool to infer ordinary differential equation representations of dynamical models. *Nonlinear Processes in Geophysics*, 26(3):143–162, jul 2019.

[55] Tapio Schneider, Shiwei Lan, Andrew Stuart, and João Teixeira. Earth System Modeling 2.0: A Blueprint for Models That Learn From Observations and Targeted High-Resolution Simulations. *Geophysical Research Letters*, 44(24):12,396–12,417, dec 2017.

[56] Andreas Raue, C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmüller, and J. Timmer. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15):1923–1929, aug 2009.

[57] Ottar N. Bjørnstad and Bryan T. Grenfell. Noisy clockwork: Time series analysis of population fluctuations in animals. *Science*, 293(5530):638–643, jul 2001.

[58] Christopher Rackauckas and Qing Nie. DifferentialEquations.jl – A performant and feature-rich ecosystem for solving differential equations in Julia. *Journal of Open Research Software*, 5, 2017.

[59] Yingbo Ma, Vaibhav Dixit, Michael J. Innes, Xingjian Guo, and Chris Rackauckas. A comparison of automatic differentiation and continuous sensitivity analysis for derivatives of differential equation solutions. In *2021 IEEE High Performance Extreme Computing Conference (HPEC)*, number 2, pages 1–9. IEEE, sep 2021.

[60] Amir Gholaminejad, Kurt Keutzer, and George Biros. ANODE: Unconditionally Accurate Memory-Efficient Gradients for Neural ODEs. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, volume 2019-Augus, pages 730–736, California, aug 2019. International Joint Conferences on Artificial Intelligence Organization.

[61] Patrick K Mogensen and Asbjørn N Riseth. Optim: A mathematical optimization package for Julia. *Journal of Open Source Software*, 3(24):615, apr 2018.

[62] Sebastian Ruder. An overview of gradient descent optimization algorithms. pages 1–14, 2016.

[63] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. pages 1–15, dec 2014.

[64] Dong C. Liu and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45(1-3):503–528, aug 1989.

[65] Alan Hastings and Thomas Powell. Chaos in a three-species food chain. *Ecology*, 72(3):896–903, jun 1991.

[66] Kevin McCann and Peter Yodzis. Nonlinear dynamics and population disappearances. *The American Naturalist*, 144(5):873–879, nov 1994.

[67] Kevin McCann and Peter Yodzis. Biological conditions for chaos in a three-species food chain. *Ecology*, 75(2):561–564, mar 1994.

[68] Aaron Klebanoff and Alan Hastings. Chaos in three species food chains. *Journal of Mathematical Biology*, 32(5):427–451, may 1994.

[69] Kevin McCann and Alan Hastings. Re–evaluating the omnivory–stability relationship in food webs. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 264(1385):1249–1254, aug 1997.

[70] David M. Post, M. Elizabeth Conners, and Debra S. Goldberg. Prey preference by a top predator and the stability of linked food chains. *Ecology*, 81(1):8–14, 2000.

[71] Charles T. Perretti, George Sugihara, and Stephan B. Munch. Nonparametric forecasting outperforms parametric methods for a simulated multispecies system. *Ecology*, 94(4):794–800, apr 2013.

[72] Ruiwen Dong, Christian Goodbrake, Heather A Harrington, and Gleb Pogudin. Differential elimination for dynamical models via projections with applications to structural identifiability. pages 1–37, nov 2021.

[73] N. M. Mangan, J. N. Kutz, S. L. Brunton, and J. L. Proctor. Model selection for dynamical systems via sparse regression and information criteria. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 473(2204):20170009, aug 2017.

[74] David R. Anderson Kenneth P. Burnham and Model. *Model Selection and Multimodel Inference*. Springer New York, New York, NY, 2002.

[75] Claus O. Wilke, Jia Lan Wang, Charles Ofria, Richard E. Lenski, and Christoph Adami. Evolution of digital organisms at high mutation rates leads to survival of the flattest. *Nature*, 412(6844):331–333, jul 2001.

[76] Maria Rodriguez-Fernandez, Pedro Mendes, and Julio R. Banga. A hybrid approach for efficient and robust parameter estimation in biochemical pathways. *Biosystems*, 83(2-3):248–265, feb 2006.

[77] Maria Dornelas, Laura H. Antão, Faye Moyes, Amanda E. Bates, Anne E. Magurran, Dušan Adam, Asem A. Akhmetzhanova, Ward Appeltans, José Manuel Arcos, Haley Arnold, Narayanan Ayyappan, Gal Badihi, Andrew H. Baird, Miguel Barbosa, Tiago Egydio Barreto, Claus Bässler, Alecia Bellgrove, Jonathan Belmaker, Lisandro Benedetti-Cecchi, Brian J. Bett, Anne D. Bjorkman, Magdalena Błażewicz, Shane A. Blowes, Christopher P. Bloch, Timothy C. Bonebrake, Susan Boyd, Matt Bradford, Andrew J.

Brooks, James H. Brown, Helge Bruelheide, Phaedra Budy, Fernando Carvalho, Edward Castañeda-Moya, Chaolun Allen Chen, John F. Chamblee, Tory J. Chase, Laura Siegwart Collier, Sharon K. Collinge, Richard Condit, Elisabeth J. Cooper, J. Hans C. Cornelissen, Unai Cotano, Shannan Kyle Crow, Gabriella Damasceno, Claire H. Davies, Robert A. Davis, Frank P. Day, Steven Degraer, Tim S. Doherty, Timothy E. Dunn, Giselda Durigan, J. Emmett Duffy, Dor Edelist, Graham J. Edgar, Robin Elahi, Sarah C. Elmendorf, Anders Enemar, S. K. Morgan Ernest, Rubén Escribano, Marc Estiarte, Brian S. Evans, Tung-Yung Fan, Fabiano Turini Farah, Luiz Loureiro Fernandes, Fábio Z. Farneda, Alessandra Fidelis, Robert Fitt, Anna Maria Fosaa, Geraldo Antonio Daher Correa Franco, Grace E. Frank, William R. Fraser, Hernando García, Roberto Cazzolla Gatti, Or Givan, Elizabeth Gorgone-Barbosa, William A. Gould, Corinna Gries, Gary D. Grossman, Julio R. Gutierréz, Stephen Hale, Mark E. Harmon, John Harte, Gary Haskins, Donald L. Henshaw, Luise Hermanutz, Pamela Hidalgo, Pedro Higuchi, Andrew Hoey, Gert Van Hoey, Annika Hofgaard, Kristen Holeck, Robert D. Hollister, Richard Holmes, Mia Hoogenboom, Chih-hao Hsieh, Stephen P. Hubbell, Falk Huettmann, Christine L. Huffard, Allen H. Hurlbert, Natália Macedo Ivanauskas, David Janík, Ute Jandt, Anna Jażdżewska, Tore Johannessen, Jill Johnstone, Julia Jones, Faith A. M. Jones, Jungwon Kang, Tasrif Kartawijaya, Erin C. Keeley, Douglas A. Kelt, Rebecca Kinnear, Kari Klanderud, Halvor Knutsen, Christopher C. Koenig, Alessandra R. Kortz, Kamil Král, Linda A. Kuhnz, Chao-Yang Kuo, David J. Kushner, Claire Laguionie-Marchais, Lesley T. Lancaster, Cheol Min Lee, Jonathan S. Lefcheck, Esther Lévesque, David Lightfoot, Francisco Lloret, John D. Lloyd, Adrià López-Baucells, Maite Louzao, Joshua S. Madin, Borgþór Magnússon, Shahar Malamud, Iain Matthews, Kent P. McFarland, Brian McGill, Diane McKnight, William O. McLarney, Jason Meador, Peter L. Meserve, Daniel J. Metcalfe, Christoph F. J. Meyer, Anders Michelsen, Nataliya Milchakova, Tom Moens, Even Moland, Jon Moore, Carolina Mathias Moreira, Jörg Müller, Grace Murphy, Isla H. Myers-Smith, Randall W. Myster, Andrew Naumov, Francis Neat, James A. Nelson, Michael Paul Nelson, Stephen F. Newton, Natalia Norden, Jeffrey C. Oliver, Esben M. Olsen, Vladimir G. Onipchenko, Krzysztof Pabis, Robert J. Pabst, Alain Paquette, Sinta Pardede, David M. Paterson, Raphaël Pélissier, Josep Peñuelas, Alejandro Pérez-Matus, Oscar Pizarro, Francesco Pomati, Eric Post, Herbert H. T. Prins, John C. Priscu, Pieter Provoost, Kathleen L. Prudic, Erkki Pulliainen, B. R. Ramesh, Olivia Mendivil Ramos, Andrew Rassweiler, Jose Eduardo Rebelo, Daniel C. Reed, Peter B. Reich, Suzanne M. Remillard, Anthony J. Richardson, J. Paul Richardson, Itai van Rijn, Ricardo Rocha, Victor H. Rivera-Monroy, Christian Rixen, Kevin P. Robinson, Ricardo Ribeiro Rodrigues, Denise de Cerqueira Rossa-Feres, Lars Rudstam, Henry Ruhl, Catalina S. Ruz, Erica M. Sampaio, Nancy Rybicki, Andrew Rypel, Sofia Sal, Beatriz Salgado, Flavio A. M. San-

tos, Ana Paula Savassi-Coutinho, Sara Scanga, Jochen Schmidt, Robert Schooley, Fakhrizal Setiawan, Kwang-Tsao Shao, Gaius R. Shaver, Sally Sherman, Thomas W. Sherry, Jacek Siciński, Caya Sievers, Ana Carolina da Silva, Fernando Rodrigues da Silva, Fabio L. Silveira, Jasper Slingsby, Tracey Smart, Sara J. Snell, Nadejda A. Soudzilovskaia, Gabriel B. G. Souza, Flaviana Maluf Souza, Vinícius Castro Souza, Christopher D. Stallings, Rowan Stanforth, Emily H. Stanley, José Mauro Sterza, Maarten Stevens, Rick Stuart-Smith, Yzel Rondon Suarez, Sarah Supp, Jorge Yoshio Tamashiro, Sukmaraharja Tarigan, Gary P. Thiede, Simon Thorn, Anne Tolvanen, Maria Teresa Zugliani Toniato, Ørjan Totland, Robert R. Twilley, Gediminas Vaitkus, Nelson Valdivia, Martha Isabel Vallejo, Thomas J. Valone, Carl Van Colen, Jan Vanaverbeke, Fabio Venturoli, Hans M. Verheye, Marcelo Vianna, Rui P. Vieira, Tomáš Vrška, Con Quang Vu, Lien Van Vu, Robert B. Waide, Conor Waldock, Dave Watts, Sara Webb, Tomasz Wesołowski, Ethan P. White, Claire E. Widdicombe, Dustin Wilgers, Richard Williams, Stefan B. Williams, Mark Williamson, Michael R. Willig, Trevor J. Willis, Sonja Wipf, Kerry D. Woods, Eric J. Woehler, Kyle Zawada, Michael L. Zettler, and Thomas Hickler. BioTIME: A database of biodiversity time series for the Anthropocene. *Global Ecology and Biogeography*, 27(7):760–786, jul 2018.

[78] M. Raissi, P. Perdikaris, and G.E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, feb 2019.

[79] Alireza Yazdani, Lu Lu, Maziar Raissi, and George Em Karniadakis. Systems biology informed deep learning for inferring parameters and hidden dynamics. *PLOS Computational Biology*, 16(11):e1007575, nov 2020.

[80] J. O. Ramsay, G. Hooker, D. Campbell, and J. Cao. Parameter estimation for differential equations: a generalized smoothing approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(5):741–796, nov 2007.

[81] Jiguo Cao, Gregor F. Fussmann, and James O. Ramsay. Estimating a predator-prey dynamical model with the parameter cascades method. *Biometrics*, 64(3):959–967, sep 2008.

[82] M. L. Pinsky, B. Worm, M. J. Fogarty, J. L. Sarmiento, and S. A. Levin. Marine taxa track local climate velocities. *Science*, 341(6151):1239–1242, sep 2013.

[83] Michael T. Burrows, Amanda E. Bates, Mark J. Costello, Martin Edwards, Graham J. Edgar, Clive J. Fox, Benjamin S. Halpern, Jan G. Hiddink, Malin L. Pinsky, Ryan D. Batt, Jorge García Molinos, Benjamin L.

Payne, David S. Schoeman, Rick D. Stuart-Smith, and Elvira S. Poloczanska. Ocean community warming responses explained by thermal affinities and temperature gradients. *Nature Climate Change*, 9(12):959–963, dec 2019.

[84] Chih-hao Hsieh, Christian Anderson, and George Sugihara. Extending nonlinear analysis to short ecological time series. *The American Naturalist*, 171(1):71–80, jan 2008.

[85] Adam Thomas Clark, Hao Ye, Forest Isbell, Ethan R. Deyle, Jane Cowles, G. David Tilman, George Sugihara, and B. D. Inouye. Spatial convergent cross mapping to detect causal relationships from short time series. *Ecology*, 96(5):1174–1181, 2015.

[86] H.T. Banks, J.E. Banks, R. Bommarco, A. Curtsdotter, T. Jonsson, and A.N. Laubmeier. Parameter estimation for an allometric food web model. *International Journal of Pure and Apllied Mathematics*, 114(1):143–160, apr 2017.

[87] A. N. Laubmeier, Kate Wootton, J. E. Banks, Riccardo Bommarco, Alva Curtsdotter, Tomas Jonsson, Tomas Roslin, and H. T. Banks. From theory to experimental design—Quantifying a trait-based theory of predator-prey dynamics. *PLoS ONE*, 13(4):e0195919, apr 2018.

[88] Mark A. Beaumont. Approximate Bayesian computation in evolution and ecology. *Annual Review of Ecology, Evolution, and Systematics*, 41(1):379–406, 2010.

[89] Markus Pahlow, Alain F. Vézina, Benoit Casault, Heidi Maass, Louise Malloch, Daniel G. Wright, and Youyu Lu. Adaptive model of plankton dynamics for the North Atlantic. *Progress in Oceanography*, 76(2):151–191, 2008.

[90] Léon Bottou. Stochastic gradient descent tricks. In *Neural networks: Tricks of the trade*, pages 421–436. Springer, 2012.

[91] Krista M. Ruppert, Richard J. Kline, and Md Saydur Rahman. Past, present, and future perspectives of environmental DNA (eDNA) metabarcoding: A systematic review in methods, monitoring, and applications of global eDNA. *Global Ecology and Conservation*, 17:e00547, jan 2019.

[92] Walter Jetz, Melodie A. McGeoch, Robert Guralnick, Simon Ferrier, Jan Beck, Mark J. Costello, Miguel Fernandez, Gary N. Geller, Petr Keil, Cory Merow, Carsten Meyer, Frank E. Muller-Karger, Henrique M. Pereira, Eugenie C. Regan, Dirk S. Schmeller, and Eren Turak. Essential biodiversity variables for mapping and monitoring species populations. *Nature Ecology and Evolution*, 3(4):539–551, 2019.

[93] T. Mitchell Aide, Carlos Corrada-Bravo, Marconi Campos-Cerqueira, Carlos Milan, Giovany Vega, and Rafael Alvarez. Real-time bioacoustics monitoring and automated species identification. *PeerJ*, 1(1):e103, jul 2013.

[94] GBIF: The Global Biodiversity Information Facility. What is GBIF? 2022.

[95] Steven H Strogatz. *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering*. CRC press, 2018.

[96] Jürgen Jost. *Dynamical systems: examples of complex behaviour*. Springer Science & Business Media, 2005.

S1 Supplementary Information

We show that the loss function $L_{\mathcal{M}}(\theta)$ in the main manuscript in Eq. (6) is ill-behaved for models with complex dynamics when the time horizon is large. We proceed by first analysing the dynamics of models with complex dynamics showing chaotic behaviour or limit cycles, and approximate the divergence of perturbed dynamical trajectories. We then show that the divergence in dynamics translates into a loss function whose surface is rugged in most of the parameter space, and that the gradient of the loss function around the true parameters becomes exponentially steeper with time. We conclude by formally discussing how the proposed mini-batch method regularizes the ill-behaviour of the loss function.

S1.1 Dynamics under perturbations

S1.1.1 Perturbed initial conditions

Consider the trajectory of the state variables

$$\begin{aligned} x(t) &= \mathcal{M}(t, p, x_0) \\ &= \int_0^t f(s, x(s), p) ds + x_0 \end{aligned} \tag{S1}$$

and consider the perturbed trajectory

$$x_{\delta x_0}(t) = \mathcal{M}(t, p, x_0 + \delta x_0) \tag{S2}$$

whose initial conditions (ICs) x_0 are perturbed by δx_0. Assuming that the system is chaotic and that δx_0 is small, the distance between the perturbed trajectory and the original one grows as

$$\|x(t) - x_{\delta x_0}(t)\| \sim e^{\lambda t} \delta x_0 \tag{S3}$$

where λ is the largest Lyapunov exponent of the system [95]. After enough time, the trajectories diverge so much that they effectively become independent samples of the phasespace: the trajectories forget their ICs, and ergodic theory ensures that the positions of $x(t)$ and $x_{\delta x_0}(t)$ are better described by a random variable $\mathcal{R}$ with probability density given by the density of orbits in the chaotic attractor, the so-called invariant measure of the

chaotic attractor [96]. The distance between $x(t)$ and $x_{\delta x_0}(t)$ can therefore be described as

$$\|x(t) - x_{\delta x_0}(t)\| \sim \|\mathcal{R}_1 - \mathcal{R}_2\| \tag{S4}$$

Considering the observation function h, it follows that

$$\begin{aligned} \|h(x(t)) - h(x_{\delta x_0}(t))\| &\sim \left.\frac{\partial h}{\partial x}\right|_{x(t)} e^{\lambda t}\delta x_0 \text{ for } t \ll \frac{1}{\lambda} \\ \|h(x(t)) - h(x_{\delta x_0}(t))\| &\sim \|h(\mathcal{R}_1) - h(\mathcal{R}_2)\| \text{ for } t \gg \frac{1}{\lambda} \end{aligned} \tag{S5}$$

(see Fig. S2 for an illustration of the divergence behaviour over time).

S1.1.2 Perturbed model parameters

Consider now a trajectory $x_{\delta p}$ with a small perturbation of the parameters δp . From Eqs. (S1) and (S2), it follows that

$$\begin{aligned} \dot{x}_{\delta p}(t) &= f(t, x_{\delta p}(t), p + \delta p) \\ &\sim f(t, x_{\delta p}(t), p) + \frac{\partial f(t, x_{\delta p}(t), p)}{\partial p}\delta p \\ &\sim f(t, x(t), p) + \frac{\partial f(t, x_{\delta p}(t), p)}{\partial p}\delta p + \frac{\partial f(t, x(t), p)}{\partial x(t)}\frac{\partial x(t)}{\partial p}\delta p \end{aligned} \tag{S6}$$

which is dominated by the first term under small values of δp, and is thus subject to chaotic dynamics. Similar to a perturbation of the ICs, the small perturbation δp generates a divergence in the dynamical trajectories that grows exponentially until they become uncorrelated. For small δp, the distance between the true and deviated trajectory can therefore be approximated as

$$\|x(t) - x_{\delta p}(t)\| \sim \int_0^t e^{\lambda(t-s)}\|x(s) - x_{\delta p}(s)\|ds \sim e^{\lambda t} g(\delta p) \tag{S7}$$

where $g(\delta p) = \|\frac{\partial f(t, x_{\delta p}(t), p)}{\partial p} + \frac{\partial f(t, x(t), p)}{\partial x(t)}\frac{\partial x(t)}{\partial p}\|$ gives the scale of the divergence between the two trajectories as a function of δp. Similar to a perturbation of the ICs, the difference in trajectories grows to the point where the trajectories become effectively independent after a long time. Hence, it follows that

$$\begin{aligned} \|h(x(t)) - h(x_{\delta p}(t))\| &\sim \left.\frac{\partial h}{\partial x}\right|_{x(t)} e^{\lambda t} g(\delta p) \text{ for } t \ll \frac{1}{\lambda} \\ \|h(x(t)) - h(x_{\delta p}(t))\| &\sim \|h(\mathcal{R}_1) - h(\mathcal{R}_2)\| \text{ for } t \gg \frac{1}{\lambda} \end{aligned} \tag{S8}$$

(see Fig. S3 for an illustration of the divergence behaviour over time).

In the following section, we call the first divergence regime the informative divergence regime, where the loss grows with the distance to the true parameters, and we call the second divergence regime the mixed divergence regime, where the loss is dominated by the random-like behaviour. Given that δp and δx_0 behave similarly, we employ θ and $\delta\theta$ to encompass both perturbations of parameters and ICs, and denote by $g(\delta\theta)$ the function that gives the scale of the divergence in trajectories for both $g(\delta p)$ and $\|\delta x_0\|$.

S1.1.3 Transition in the parameter space between the informative and the mixed regime

For a fixed time horizon t, and depending on the shape of the chaotic attractor, the magnitude of the perturbation determines the divergence regime. If the perturbation is small, the trajectories will be aligned, but for large perturbations they will effectively become two independent trajectories.

The transition between the two regimes can be studied by noting that the informative divergence should remain in the same order of magnitude as the mixed divergence. The reason is that the expected value of the squared divergence between two trajectories $x(t)$ and $x_{\delta\theta}(t)$ for large t is

$$\mathbb{E}\left[\|h\left(\mathcal{R}_1\right)-h\left(\mathcal{R}_2\right)\|^2\right]=2\mathrm{Var}\left[h\left(\mathcal{R}\right)\right]\lesssim\max\left|h(\mathcal{R})-\mathbb{E}\left[h(\mathcal{R})\right]\right|^2 \tag{S9}$$

meaning that the expected value of the squared divergence in the mixed regime is in the same order of magnitude as the maximum distance within the phasespace. On the other hand, the divergence of any two trajectories in the chaotic attractor cannot be larger than the maximum distance between two points in the attractor, which is itself bounded through the triangle inequality as $2\max\left|h(\mathcal{R})-\mathbb{E}\left[h(\mathcal{R})\right]\right|$.

Since the growth of the informative regime has to remain in the same order of magnitude as the mixed regime, at the regime transition we must have

$$e^{\lambda t}g(\delta\theta)\sim\mathbb{E}\left[\|h\left(\mathcal{R}\right)\|\right] \tag{S10}$$

Equation (S10) implies that, for a given time horizon t, the magnitude of the critical perturbation $\delta\theta^*$ associated with the regime transition satisfies

$$\|\delta\theta^*\|\sim e^{-\lambda t} \tag{S11}$$

S1.1.4 Limit cycles

While in the section above a chaotic system was assumed to provide an approximation for the divergence of the trajectories, a similar approximation applies for systems characterized by limit cycles. Considering a system $x(t)$ with a limit cycle characterized by the phase ωt with frequency ω, i.e.

$$x(t) = \mathfrak{f}(\phi(t)) = \mathfrak{f}(\omega t) \mod 2\pi \tag{S12}$$

a perturbation of the parameters δp might lead to a perturbed frequency $\delta\omega$, further leading to a difference in phases

$$\delta\phi(t) = \phi(t) - \phi_{\delta\omega}(t) = \delta\omega t \mod 2\pi \tag{S13}$$

For $t \lesssim \frac{1}{\delta\omega}$, $\delta\phi(t)$ grows linearly with $\delta\omega$, but once $t \gg \frac{1}{\delta\omega}$, the change of phase $\delta\phi(t)$ is affected by the modulo operation. As this operation is nonlinear, a small random perturbation $\delta\omega$ results in a random uniform phase over the interval $[0, 2\phi]$. For a large time horizon, $x_{\delta\omega}$ is thus uniformly spread over the circular line given by the phasespace of the dynamical system. Hence, the approximation in Eq. (S8) applies for cyclic dynamics, except that the initial divergence is linear rather than exponential, and that in contrast to Eq. (S5), a change affecting the initial position will not grow over time.

S1.2 Consequences for the shape of the loss surface

The approximation of the divergence of trajectories in Eqs. (S5) and (S8), together with the transition boundary determined by Eq. (S11), can be used to characterize the surface associated with the loss function.

Omitting the term corresponding to the priors and the variance–covariance matrix Σ_y for simplicity, the loss function presented in the main text is expressed as

$$\begin{aligned} L_{\mathcal{M}}(\theta) &= \frac{1}{K}\sum_{k=1}^{K} ||y_k - h(\mathcal{M}(t_k, \theta))||^2 \\ &= \frac{1}{K}\sum_{k=1}^{K} ||h(\tilde{x}(t_k)) + \epsilon(t_k) - h(\mathcal{M}(t_k, \theta))||^2 \end{aligned} \tag{S14}$$

where the parameter vector θ is decomposed into the model parameter vector p and the ICs x_0, y_k correspond to the observations, $\tilde{x}$ corresponds to the true trajectory and $\epsilon(t)$ is the observational noise. As the noise is independent of the dynamics, it is uncorrelated with $h(\tilde{x}(t_k)) - \mathcal{M}(t_k, \theta)$, meaning that the loss can be split in

expectation

$$\mathbb{E}\left[L_{\mathcal{M}}(\theta)\right] = \frac{1}{K}\sum_{k=1}^{K} ||h(\tilde{x}(t_k)) - h(\mathcal{M}(t_k,\theta))||^2 + \mathrm{Var}\left[\epsilon\right] \tag{S15}$$

where the noise term ($\mathrm{Var}\left[\epsilon\right]$) is independent from the parameters θ. Assuming that $\theta = \tilde{\theta} + \delta\theta$, where $\tilde{\theta}$ correspond to the true parameters and ICs, every term in the sum corresponds to a squared distance between the true trajectory $\tilde{x}(t)$ and a perturbed trajectory $\tilde{x}_{\delta\theta}(t)$. Using Eqs. (S5) and (S8) to (S10) we obtain the loss function approximation

$$L_{\mathcal{M}}(\theta) \sim \sum_{k=1}^{\min\{K,K^*\}} e^{2\lambda t_k} g^2(\delta\theta) + \sum_{k=\min\{K,K^*\}+1}^{K} \mathrm{Var}\left[h(\mathcal{R})\right] + \mathrm{Var}\left[\epsilon\right] \tag{S16}$$

where K^* is the observation index corresponding to the time horizon t_{K^*}, where the transition between the informative and the mixed regime happens for the perturbation $\delta\theta$, obtained from Eq. (S11).

The distribution of the observation times and the magnitude of the perturbation $\delta\theta$ determine whether the loss is dominated by the informative or by the mixed divergence regime. For a fixed perturbation $\delta\theta$, assuming that the observations y_k are uniformly distributed over the time interval $[0,t]$, Eq. (S11) yields for $t \ll \frac{\log(|\delta\theta|)}{\lambda}$ that the loss $L_{\mathcal{M}}$ is dominated by the informative divergence regime, whereas if $t \gg \frac{\log(|\delta\theta|)}{\lambda}$ the loss is dominated by the mixed divergences. In the region where the loss is dominated by mixed divergences, the loss $L_{\mathcal{M}}$ has an expected value of order $\mathcal{O}\left(\mathrm{Var}\left[h(\mathcal{R})\right] + \mathrm{Var}\left[\epsilon\right]\right)$ and does not grow monotonically with $\delta\theta$. It corresponds to a "random-like" surface populated with local minima, and is consequently characterized by an uninformative gradient preventing local optimizers from converging to the true parameters $\tilde{\theta}$ (see Fig. 1B, orange dashed curve). On the other hand, in the region where the divergence in trajectories belongs to the informative regime, the loss is convex and grows with $\delta\theta$. Its associated gradient $\nabla_\theta L_{\mathcal{M}}(\theta) \sim \sum_{k=1}^{K} e^{2\lambda t_k} \nabla_\theta g^2(\delta\theta)$ consequently contains relevant information for the use of variational optimizers, but the loss surface becomes exponentially steeper as the time horizon increases. As shown by the green dotted curve in Fig. 1B, this large gradient in the vicinity of the optimal parameters is likely to lead gradient-based optimizers to overshoot and not converge to the true parameters $\tilde{\theta}$. Eq. (S11) further indicates that the volume of the region in the parameter space where the loss is informative shrinks exponentially as the time horizon increases, implying that for large time horizons, the uninformative region is predominant.

S1.3 Regularizing the loss surface with mini-batches

To prevent the situation with a mixed divergence regime and to decrease the gradient in a vicinity of the true parameters, we reformulate the loss function as the average of loss functions defined over mini-batches of short time horizons. In the following section, we compare the properties of the naive loss function $L_{\mathcal{M}}$ in Eq. (S15) with the mini-batch loss function $L^{\star}_{\mathcal{M}}$ proposed in Eq. (8) in the main manuscript, and further discuss the limitations of the method in the presence of noise.

Omitting the term corresponding to the priors and ICs and the noise shape Σ_y, and assuming a single time series for simplicity, the mini-batch loss function presented in the main text can be expressed as

$$L^{\star}_{\mathcal{M}}(\theta) = \frac{1}{M}\sum_{m=0}^{M-1}\frac{1}{K}\sum_{k=1}^{K^{(m)}} \| y_{k+mK/M} - h\left(\mathcal{M}(t_{k+mK/M}, p, x_0^{(m)})\right) \|^2 \tag{S17}$$

where M is the number of mini-batches, $x_0^{(m)}$ corresponds to the ICs for mini-batch m inferred at time $t_{mK/M}$, and $K^{(m)} = \begin{cases} K/M+1 & : m < M \\ K/M & : m = M \end{cases}$ is the number of data points in the mth batch. The loss function can be split in expectation as

$$\mathbb{E}\left[L^{\star}_{\mathcal{M}}(\theta)\right] = \frac{1}{M}\sum_{m=0}^{M-1}\frac{1}{K}\sum_{k=1}^{K^{(m)}} \| h(\tilde{x}(t_{k+mK/M})) - \mathcal{M}(t_{k+mK/M}, p, x_0^{(m)})\|^2 + \mathrm{Var}\left[\epsilon\right] \tag{S18}$$

Assuming that $x_0^{(m)} = \tilde{x}(t_{mK/M}) + \delta x_0^{(m)}$ when $\delta x_0^{(m)}$ is small, and assuming that the observation times are regularly spaced so that $t_{k+1} - t_k = \Delta t$, the time length of simulated trajectories (i.e. the time elapsed between the time when the perturbation is applied and the end time of the simulated trajectory) is divided by the number of mini-batches, in comparison to the time length of the simulated trajectory in Eq. (S15) ($t = t_{K/M+1} - t_0 = K/M\Delta t$ in Eq. (S18), in comparison to $t = K\Delta t$ in Eq. (S15)). Further assuming that $K\Delta t >> \frac{\log(|\delta\theta|)}{\lambda}$ and choosing the number of mini-batches M so that $K/M\Delta t << \frac{\log(|\delta\theta|)}{\lambda}$, we apply the approximation Eq. (S16), which leads to

$$\begin{aligned} L_{\mathcal{M}}(\theta) &\sim \mathrm{Var}\left[\epsilon\right] + \mathrm{Var}\left[h\left(\mathcal{R}\right)\right] \\ L^{\star}_{\mathcal{M}}(\theta) &\sim \mathrm{Var}\left[\epsilon\right] + g^2(\delta\theta)\sum_{m=0}^{M-1} e^{2\lambda(K/M\Delta t)} \end{aligned} \tag{S19}$$

While $L_{\mathcal{M}}$ is dominated by the mixed regime, $L^{\star}_{\mathcal{M}}$ is dominated by the informative regime because the simulation

time remains small, permitting the successful use of variational optimizers.

The number of mini-batches M should be determined by considering the dynamical behaviour of the system and the level of noise in the observation, because a large number of mini-batches M smooths out the loss surface but also entails more sensitivity to the level of noise. Indeed, in $L_{\mathcal{M}}(\theta)^\star$ the relative effect of the second term corresponding to the observational noise ϵ increases when the number of mini-batches M increases. The value of M should therefore be chosen wisely to balance the benefits of mini-batches, i.e. widening the region of the parameter space where $L_{\mathcal{M}}$ is well behaved and reducing the overshooting problem, and their cons, i.e. their tendency to increase the importance of noise.

S2 Three-compartment food-web models

S2.1 Reference food-web model

We used the three-species chaotic food-web model from [65], formulated as

$$\begin{aligned}
\tfrac{d}{dt}R &= R(1-R) - x_C y_C \frac{CR}{R+R_0} \\
\tfrac{d}{dt}C &= x_C C\left[-1 + x_C\frac{R}{R+R_0}\right] - x_P y_P \frac{PC}{C+C_0} \\
\tfrac{d}{dt}P &= x_P P\left[-1 + y_P\frac{C}{C+C_0}\right],
\end{aligned} \tag{S20}$$

with the biologically realistic parameter values $x_C = 0.4$, $0.071 \leq x_P \leq 0.225$, $y_C = 2.01$, $y_P = 5$, $R_0 = 0.16129$, and $C_0 = 0.5$ [67]. The dynamics of the system are chaotic for this set of parameter values.

S2.2 Omnivory variant food-web model

We used the three-species food-web model from [69], formulated as

$$\begin{aligned}
\tfrac{d}{dt}R &= R(1-R) - x_C x_C \frac{CR}{R+R_0} - \omega x_P y_{PR}\frac{PR}{R_{02}+(1-\omega)C+\omega R} \\
\tfrac{d}{dt}C &= x_C C\left[-1 + x_C\frac{R}{R+R_0}\right] - (1-\omega)x_P y_{PC}\frac{PC}{\omega R+(1-\omega)C+C_0} \\
\tfrac{d}{dt}P &= x_P P\left[-1 + (1-\omega)x_P y_{PC}\frac{C}{\omega R+(1-\omega)C+C_0} + \omega x_P y_{PR}\frac{R}{\omega R+(1-\omega)C+R_{02}}\right]
\end{aligned} \tag{S21}$$

with the biologically realistic parameter values $x_C = 0.4$, $x_P = 0.08$, $y_C = 2.009$, $y_{PR} = 2$, $y_{PC} = 5$, $R_0 = 0.16129$, $C_0 = 0.5$, and $0 \leq \omega \leq 0.5$. For this set of parameter values, the dynamics of the system are chaotic

for $\omega \lesssim 0.20$, consist of a limit cycle for $0.20 \lesssim \omega \lesssim 0.35$, and consist of dampened oscillations for $0.35 \lesssim \omega$.

S3 Supplementary Figures

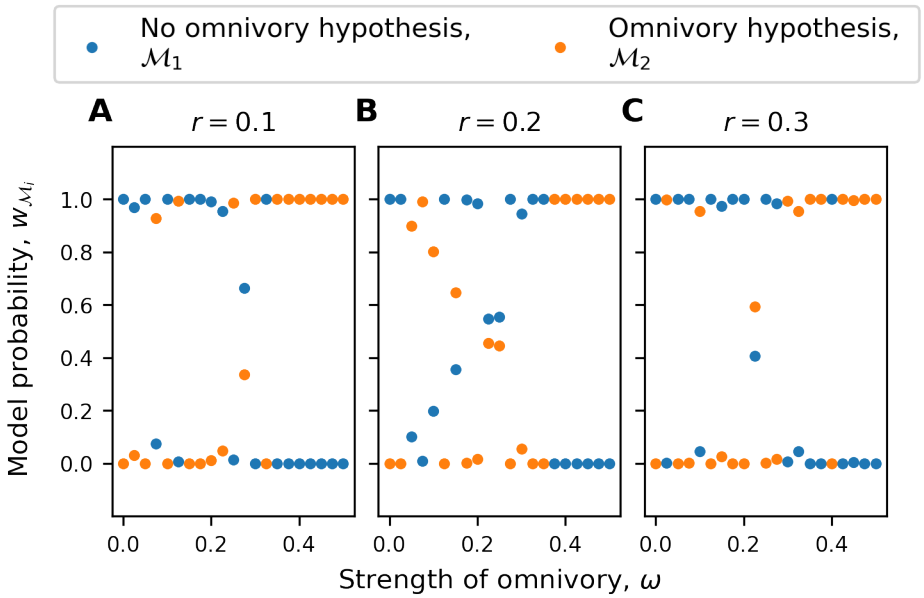

Figure S1: **Performance of the ML framework in supporting the predator omnivory hypothesis in a food web for the partial observation setting**. In **A**, **B** and **C** where $r = 0.1, 0.2, 0.3$, the lack of data prevents the correct estimation of the omnivory variant model parameters, leading model $\mathcal{M}_1$ to be supported for a wider range of ω in contrast to the complete observation setting (Fig. 5A).

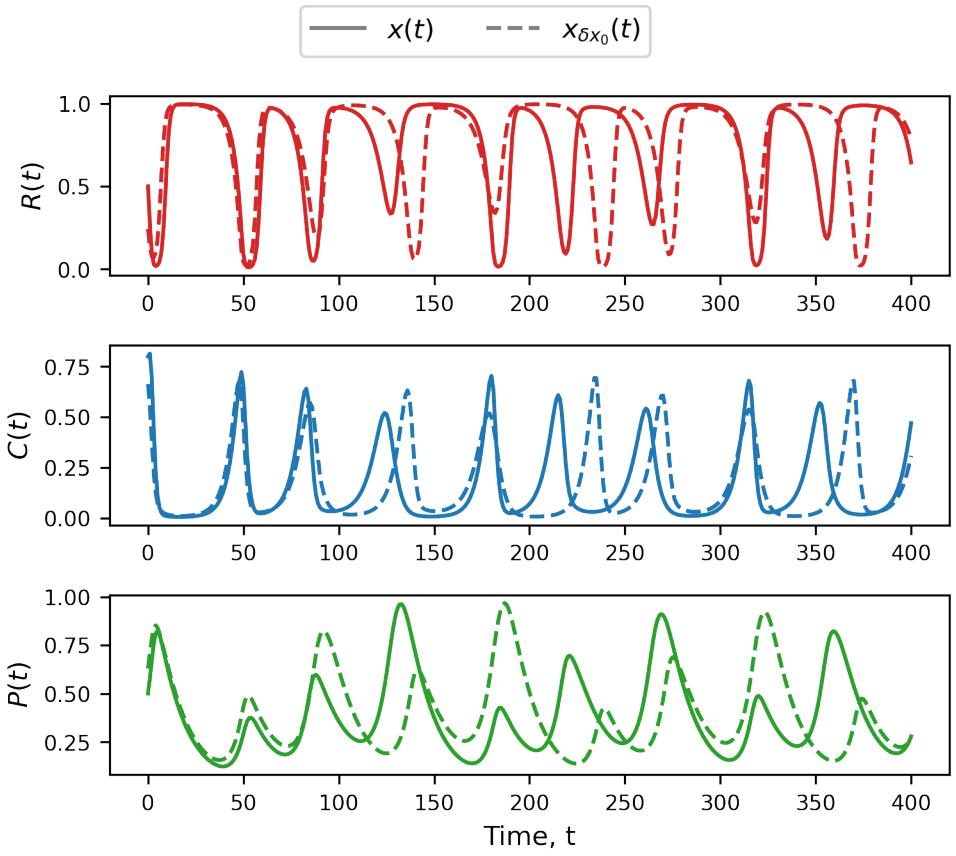

Figure S2: **Divergence between the trajectory $x(t)$ and a perturbed trajectory $x_{\delta x_0}(t)$, obtained from the reference food-web model from [65] and detailed in Section S2**. For $t \lesssim 100$, $x(t)$ and $x_{\delta x_0}(t)$ are correlated and the divergence regime is informative, but for $t \gtrsim 100$ the trajectories become essentially uncorrelated, corresponding to the mixed divergence regime.

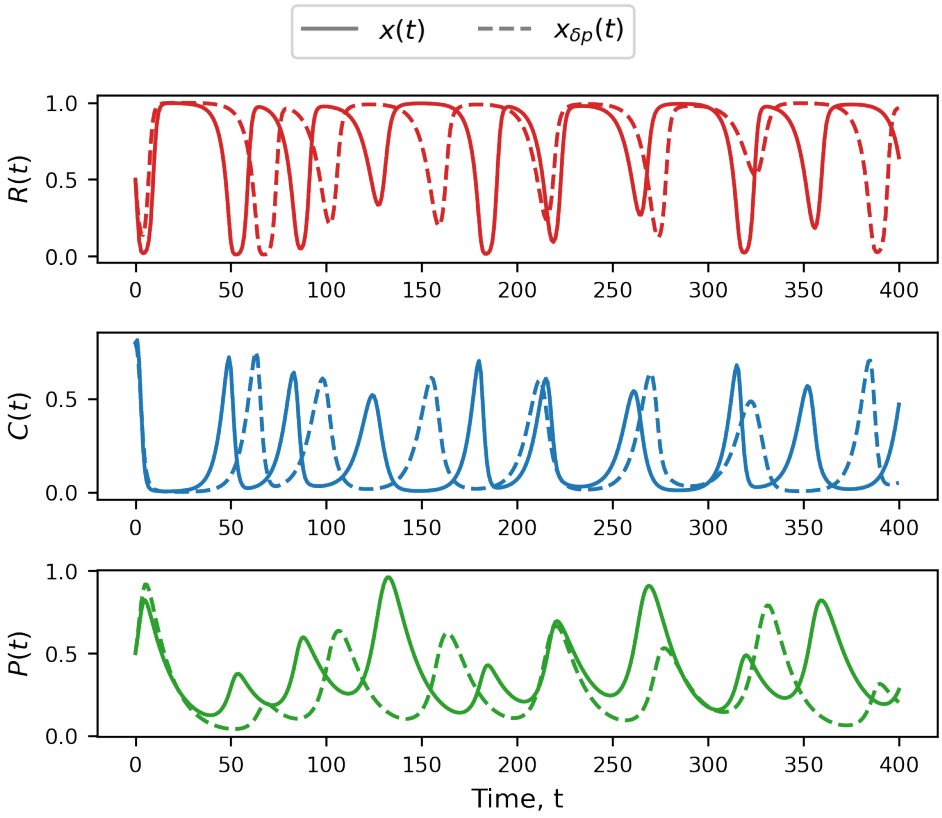

Figure S3: **Divergence between the trajectory $x(t)$ and a perturbed trajectory $x_{\delta p}(t)$, obtained from the reference food-web model from [65] and detailed in Section S2**. For $t \lesssim 40$, $x(t)$ and $x_{\delta p}(t)$ are correlated and the divergence regime is informative, but for $t \gtrsim 40$ the trajectories become essentially uncorrelated, corresponding to the mixed divergence regime.

S4 Supplementary Tables

Setting	Median simulation time	Mean simulation time	Std. simulation time
Complete observations	35.9977436	39.5980174	20.5789433
Partial observations	34.1293998	39.1896534	21.8191709

Table S1: Simulation time for the complete and partial observation settings.