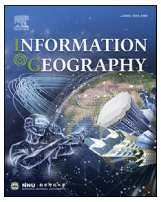
Information Geography 1 (2025) 100001

Contents lists available at ScienceDirect

Information Geography

journal homepage: www.sciencedirect.com/journal/information-geography



Research Article

A mathematical comparison of data assimilation and machine learning in earth system state estimation from a Bayesian inference viewpoint

Xin Li [a,*], Feng Liu [b,**]

[a] National Tibetan Plateau Data Center, State Key Laboratory of Tibetan Plateau Earth System, Environment and Resources, Institute of Tibetan Plateau Research, Chinese Academy of Sciences, Beijing 100101, China

[b] Key Laboratory of Cryospheric Science and Frozen Soil Engineering, Key Laboratory of Remote Sensing of Gansu Province, Heihe Remote Sensing Experimental Research Station, Northwest Institute of Eco-Environment and Resources, Chinese Academy of Sciences, Lanzhou 730000, China

ARTICLE INFO

Keywords:
Data assimilation
Machine learning
Bayesian inference
Cost function

ABSTRACT

Data Assimilation (DA) and Machine Learning (ML) possess complementary strengths that are vital to improving state estimation and prediction in Earth system science. However, comprehensive comparisons of their methodologies are rare and complex. This study explores DA and ML through a mathematical framework based on Bayesian inference. We establish a unified framework to address state estimation and prediction, deriving cost functions of DA and ML using maximum a posteriori estimation. While DA and ML share similarities in cost function structures, they diverge in principles and applications. DA is rooted in physical models, while ML excels with large datasets. Despite their similarities in motivations, distinctions arise in equations, algorithms, cost functions, parameters, uncertainty handling, and data requirements. Integrating DA's theoretical robustness with ML's data leveraging capabilities is essential for advancing state estimation and prediction. DA can bolster ML's interpretability, while ML can boost DA's operational efficiency. This paper contributes to the discourse on how best to integrate these methodologies for improving state estimation and prediction in Earth system science.

1. Introduction

Advancing accuracy and reliability of state estimation and enhancing predictability remains a cornerstone challenge in Earth system science. Addressing this challenge hinges on leveraging two pivotal methodologies: data assimilation (DA) and machine learning (ML). DA aims to enhance the predictability and reduce uncertainties in Earth system dynamics by blending dynamical models with observational data (Gettelman et al., 2022; Li et al., 2020, 2024; McLaughlin, 1995). Conversely, ML makes the prediction by optimizing model parameters, identifying patterns and discerning underlying system behaviors, primarily utilizing empirical data. Although ML is currently less dependent on prior knowledge, its predictive ability can be further improved by integrating a more nuanced understanding of Earth system dynamics (Brajard et al., 2021; Carrassi et al., 2018).

Both DA and ML are undergoing rapid development and have their distinct advantages. DA excels in modeling well-described Earth system dynamics governed by mathematical equations and physical laws (Carrassi et al., 2018). ML is highly effective for modeling complex systems that lack governing equations or a comprehensive understanding of the Earth system dynamics (Ghahramani, 2015). The fusion of increased computational resources and expansive datasets (Li et al., 2023) has propelled ML forward, culminating in successful applications like ECMWF's recent advancements (Geer, 2021), innovative nowcasting with deep generative models (Ravuri et al., 2021), and meteorological forecasting through NowcastNet (Zhang et al., 2023), GraphCast (Lam et al., 2023), Pangu-Weather (Bi et al., 2023), and NeuralGCM (Kochkov et al., 2024).

DA and ML have been increasingly recognized as complementary, with growing interest in their integration for both scientific research and practical applications in Earth system science (Course and Nair, 2023; Gao et al., 2024; Penny et al., 2022). This intersection, which combines DA's theoretical robustness based on Bayesian inference with ML's optimization and statistics, provides fertile ground for novel insights and enhanced model-data utilization. However, comparisons of these methodologies, especially within the realm of state estimation and predictive problem-solving in Earth system, are scarce and thus present a nuanced challenge. This necessitates a focused mathematical comparison between

* Corresponding author.
** Corresponding author.
E-mail addresses: xinli@itpcas.ac.cn (X. Li), liufeng@lzb.ac.cn (F. Liu).

https://doi.org/10.1016/j.infgeo.2025.100001
Received 30 October 2024; Received in revised form 4 February 2025; Accepted 12 February 2025
Available online 19 February 2025

DA and ML to establish a solid theoretical ground and identify where integration can be most effective.

This paper is organized as follows: After providing a rationale for the mathematical comparison of DA and ML in this section, section 2 explores the distinctions between DA and ML in Earth system state estimation through the lens of Bayesian inference. Section 3 investigates the potential for cross-learning between DA and ML. Lastly, the paper calls for future research into combining DA and ML to enhance the harmonization of theory and data in Earth system prediction.

2. Mathematical comparison of DA and ML

There are two primary statistical methodologies for exploring DA and ML: frequentist estimators, with maximum likelihood estimation as a notable example, and Bayesian inference. In this study, the latter is preferred for its capacity to (1) integrate prior knowledge and thus accommodate both physics-based and data-driven models; and (2) facilitate the updating of posterior knowledge by incorporating different types of observational data and quantifying uncertainties associated with models and data (Wikle and Berliner, 2007).

2.1. Problem statement

This section defines the problem which involves modeling sequential states and observations and using these models to estimate system states from time 0 up to time T (the present time) (see Fig. 1) and make a prediction beyond time T till $T+h$.

States: Define an N-dimensional sequence of discrete states from times 0 to T, i.e., $\mathbf{x} = \{\mathbf{x}_t | t = 0, 1, \cdots, T;\ \ \mathbf{x}_t \in \mathbb{R}^{1\times N}\}$. A sub sequence of $\mathbf{x}$ can be denoted by $\mathbf{x}_{i:j}$ $(0 \le i < j \le T)$, where $\mathbf{x}_0$ is the initial state value. $\mathbf{x}_{T+h}$ represents the future states to be predicted.

Observations: Define an M-dimensional sequence of discrete observations from time 0 to T, denoted $\mathbf{y} = \{\mathbf{y}_t | t = 0, 1, \cdots, T, \mathbf{y}_t \in \mathbb{R}^{1\times M}\}$. A sub sequence of $\mathbf{y}$ can also be denoted by $\mathbf{y}_{i:j}$ $(0 \le i < j \le T)$.

Training data: $\mathcal{D} = \{(\mathbf{y}_0, \mathbf{x}_0), \ldots, (\mathbf{y}_t, \mathbf{x}_t)\}$. In ML models, $\mathbf{y}$ represents the input data, while $\mathbf{x}$ represents the targeted/labeled output data in supervised learning. Training data can be a subset of $\{(\mathbf{y}_0, \mathbf{x}_0), \ldots, (\mathbf{y}_t, \mathbf{x}_t)\}$, also denoted as $\mathcal{D}$. The data can be temporally asynchronous, such as having $\mathbf{y}_0$ correspond to $\mathbf{x}_1$, which implies estimating the system state at the next time step using observations from the previous time step. Noted that, we use the same notations in DA and ML, but they have different meanings and implications. In DA, $\mathbf{x}$ is usually spatiotemporally continuous and dense, but $\mathbf{y}$ is comparatively sparse. In ML, a lot of observational data $\mathbf{y}$ should be collected as input, but how to label these input data with reliable states $\mathbf{x}$ is challenging.

Dynamical model: A mapping $M : \mathbf{x}_{t-1} \to \mathbf{x}_t$ describes the evolution of the states over time.

Observational model: A function $H : \mathbf{x}_t \to \mathbf{y}_t$ represents the process by which states are observed.

Dynamical model parameter set: Define a P-dimensional parameter set associated with dynamical model, i.e., $\boldsymbol{\theta} \in \mathbb{R}^{1\times P}$. $\boldsymbol{\theta}$ represents a set of parameters usually with physical meanings, e.g., hydraulic conductivity, heat capacity, and other parameters. $\boldsymbol{\theta}$ is usually time invariant but could be time variant.

ML model parameter set: Define a L-dimensional parameter set $\boldsymbol{\omega}$ associated with ML model, i.e., $\boldsymbol{\omega} \in \mathbb{R}^{1\times L}$. The dimension L could be very large in many cases such as deep learning, in which $\boldsymbol{\omega}$ are the weights of a neural network. Here, $\boldsymbol{\omega}$ is assumed to be time invariant. Notably, the estimation of the hyperparameters of the ML model is not discussed in this study.

A parametric model: A function $F_{\boldsymbol{\omega}} : \mathbf{y} \to \mathbf{x}$ that best fits the training data using a specific parameter set $\boldsymbol{\omega}$. $F_{\boldsymbol{\omega}}$ is usually a ML model.

In Earth sciences, state estimation and prediction are typically derived either from established dynamical models with partially known or estimated initial conditions, or from a historical sequence of input-output data. The former, based on well-established physical theories, generally employs DA, while the latter, purely data-driven, utilizes ML methodologies. The difference of the problem statement is that DA integrates observational data with dynamical models for accurate state estimation and prediction over time, while ML focuses on recognizing data patterns to predict future states, often capturing complex relationships not explicitly modeled.

2.2. Formulation of the state estimation and prediction problem in DA and ML

From a Bayesian perspective, the estimation and prediction of system states involves combining prior information about the system with observational data to update the states, thereby integrating past knowledge and current data into a coherent framework for inference.

2.2.1. State estimation and prediction problem in DA

Following the Bayesian framework, the aim of DA is to estimate the posterior probability distribution for the system states $\mathbf{x}$ (and the parameter vector $\boldsymbol{\theta}$) by utilizing a prior system dynamical model along with the observational model, given all previous observations from time 1 up to and including time T. This posterior estimation can be expresses as

$$P(\mathbf{x}, \boldsymbol{\theta} | \mathbf{y}_{1:T}), \tag{1}$$

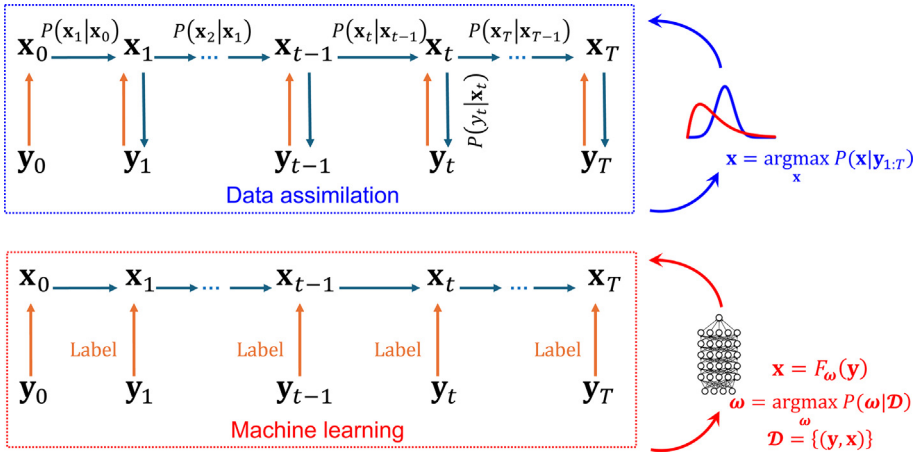


Fig. 1. Process showing how observations or data ($\mathbf{y}$) influence the system's state ($\mathbf{x}$) estimation in both data assimilation and machine learning methods, based on Bayesian theory.

According to the equation, if $\mathbf{x}$ and $\boldsymbol{\theta}$ are independent, the above expression can be written as $P(\mathbf{x}|\mathbf{y}_{1:T})P(\boldsymbol{\theta}|\mathbf{y}_{1:T})$, meaning they can be estimated separately. Most Earth system models typically use fixed parameters. Therefore, disregarding the estimation of $\boldsymbol{\theta}$ for now, the expression can be simplified as

$$P(\mathbf{x}|\mathbf{y}_{1:T}). \tag{2}$$

Noticed that, the observation at time 0 is not used in the estimation of $\mathbf{x}$ in DA, although it could be useful for initialization of the dynamical model.

The solution of equation (2) involves a prediction (or forecasting) step and an analysis step.

Prediction step: The prediction of the system states from previous time t-1 to the next time t, i.e., the state transition function, can be conceptualized using a Markov process, as:

$$P(\mathbf{x}_t|\mathbf{x}_{t-1}) = P(\mathbf{x}_t|\mathbf{x}_{0:t-1}), \tag{3}$$

which adheres to the principle that state variables, treated as discrete-time random processes, depend only on previous states.

The dynamical model in equation (3) can be explicitly represented using the mapping M defined in Section 2.1.

$$\mathbf{x}_t = M(\mathbf{x}_{t-1}) + \eta_t, \tag{4}$$

where η_t denotes the vector of cumulative model error from time t-1 to time t, which is an additive noise with arbitrary probability distribution, and $\eta_t \in \mathbb{R}^{1\times N}$. The covariance of η_t is denoted as $\mathbf{Q}_t$, and $\mathbf{Q}_t \in \mathbb{R}^{N\times N}$.

Analysis step: During the evolution of dynamical systems, there will be observations available. These observations, although incomplete in space and time, can be incorporated into a dynamical model through an observation model to improve prediction accuracy and spatiotemporal continuity. The observation model can also be expressed as a likelihood function:

$$P(\mathbf{y}_t|\mathbf{x}_t), \tag{5}$$

which can also be explicitly represented using the mapping H defined in Section 2.1.

$$\mathbf{y}_t = H(\mathbf{x}_t) + \boldsymbol{\varepsilon}_t, \tag{6}$$

where, $\boldsymbol{\varepsilon}_t$ represents the observation error vector at time t. $\boldsymbol{\varepsilon}_t$ is an additive noise with arbitrary distribution and $\boldsymbol{\varepsilon}_t \in \mathbb{R}^{1\times M}$. The covariance of $\boldsymbol{\varepsilon}_t$ is denoted as $\mathbf{R}_t$, and $\mathbf{R}_t \in \mathbb{R}^{M\times M}$.

By integrating the above equations and employing Bayesian inferences, while assuming the independence of observations, the state estimation problem in equation (2) can be sequentially expanded by combining the prediction and analysis steps as follows:

$$P(\mathbf{x}|\mathbf{y}_{1:T}) \propto P(\mathbf{x}_0)\prod_{t=1}^{T} P(\mathbf{y}_t|\mathbf{x}_t)P(\mathbf{x}_t|\mathbf{x}_{t-1}) \tag{7}$$

The prediction of future states at time $T+h$, i.e., $\mathbf{x}_{T+h}$, solely relies on the dynamical model, i.e., equations (3) and (4). However, when observational data are available in future, they will be assimilated into the dynamics to improve the estimation of future states.

2.2.2. State estimation and prediction problem in ML

Also from a Bayesian inference perspective, the state estimation problem in ML can be formulated as estimating the posterior probability distribution of the model parameter set $\boldsymbol{\omega}$ conditioned to the training data $\mathcal{D}$

$$P(\boldsymbol{\omega}|\mathcal{D}), \tag{8}$$

and using the parametric model $F_{\boldsymbol{\omega}}$ parameterized by $\boldsymbol{\omega}$ to estimate system states from time 0 up to T and make the prediction beyond time T (Goodfellow et al., 2016)

$$P(\mathbf{x},\mathbf{y}|\mathcal{D}) = \int P(\mathbf{x},\mathbf{y}|\boldsymbol{\omega})P(\boldsymbol{\omega}|\mathcal{D})d\boldsymbol{\omega}, \tag{9}$$

alternatively, this relationship can be explicitly expressed using the parametric model $F_{\boldsymbol{\omega}}$

$$\mathbf{x} = F_{\boldsymbol{\omega}}(\mathbf{y}). \tag{10}$$

To apply these equations for state estimation and prediction, the essential step is to estimate the model parameter set $\boldsymbol{\omega}$. Assuming that $\mathbf{y}_t$ and $\mathbf{x}_t$ at the same time form training pairs, equation (8) can be expanded to derive the posterior probability distribution for the parameters

$$P(\boldsymbol{\omega}|\mathcal{D}) \propto P(\mathcal{D}|\boldsymbol{\omega})P(\boldsymbol{\omega}) = P(\mathbf{x}|\mathbf{y},\boldsymbol{\omega})P(\boldsymbol{\omega}) = \prod_{t=0}^{T} P(\mathbf{x}_t|\mathbf{y}_t,\boldsymbol{\omega})P(\boldsymbol{\omega}). \tag{11}$$

Here, $P(\mathcal{D}|\boldsymbol{\omega})$ or $P(\mathbf{x}|\mathbf{y},\boldsymbol{\omega})$ represents the likelihood function given the parameter set, while $P(\boldsymbol{\omega})$ denotes the prior distribution of the parameters, providing information not contained in $\mathcal{D}$. The likelihood can be further elaborated as the product of the probabilities of the individual training pairs $(\mathbf{y}_t,\mathbf{x}_t)$ given the parameters, assuming that the training pairs are independent from each other.

2.3. Derivation of the cost/loss function

One approach to solving equations (7) and (11) is to use the maximum a posteriori (MAP) estimation algorithm, which aligns with Bayesian methods.

2.3.1. Cost function of DA

The maximum a posteriori (MAP) estimation is employed to estimate the system states that maximize the posterior probability in equation (7):

$$\mathbf{x} = \underset{\mathbf{x}}{\operatorname{argmax}}\, P(\mathbf{x}|\mathbf{y}_{1:T}) = \underset{\mathbf{x}}{\operatorname{argmax}}\, P(\mathbf{x}_0)\prod_{t=1}^{T} P(\mathbf{y}_t|\mathbf{x}_t)P(\mathbf{x}_t|\mathbf{x}_{t-1}). \tag{12}$$

Then, the logarithm is applied to both sides to transform the product into a sum, which is computationally more tractable. Additionally, the maximization problem is then converted into a minimization problem by negating the expression:

$$\begin{aligned}\underset{\mathbf{x}}{\operatorname{argmax}}\, ln\, P(\mathbf{x}|\mathbf{y}_{1:T}) =& \underset{\mathbf{x}}{\operatorname{argmax}}\left[ln\, P(\mathbf{x}_0) + \sum_{t=1}^{T} ln\, P(\mathbf{y}_t|\mathbf{x}_t) + \sum_{t=1}^{T} ln\, P(\mathbf{x}_t|\mathbf{x}_{t-1})\right] \\ =& \underset{\mathbf{x}}{\operatorname{argmin}}\left\{-\left[ln\, P(\mathbf{x}_0) + \sum_{t=1}^{T} ln\, P(\mathbf{y}_t|\mathbf{x}_t) + \sum_{t=1}^{T} ln\, P(\mathbf{x}_t|\mathbf{x}_{t-1})\right]\right\}.\end{aligned} \tag{13}$$

If the prior $P(\mathbf{x}_0)$, likelihood $P(\mathbf{y}_t|\mathbf{x}_t)$, and state transition function $P(\mathbf{x}_t|\mathbf{x}_{t-1})$ follow multivariate Normal distributions, i.e., $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{x}_b,\mathbf{B})$, $\mathbf{y}_t|\mathbf{x}_t \sim \mathcal{N}(H(\mathbf{x}_t),\mathbf{R}_t)$, and $\mathbf{x}_t|\mathbf{x}_{t-1} \sim \mathcal{N}(M(\mathbf{x}_{t-1}),\mathbf{Q}_t)$, where $\mathbf{x}_b$ is the background field of system states, and $\mathbf{B}$ is the covariance matrix of $\mathbf{x}_b$.

These prior distributions are explicitly expressed as:

$$P(\mathbf{x}_0) = (2\pi)^{-\frac{n}{2}}|\mathbf{B}|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(\mathbf{x}_0-\mathbf{x}_b)^T\mathbf{B}^{-1}(\mathbf{x}_0-\mathbf{x}_b)\right], \tag{14}$$

$$P(\mathbf{y}_t|\mathbf{x}_t) = (2\pi)^{-\frac{n}{2}}|\mathbf{R}_t|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathbf{y}_t - H_t(\mathbf{x}_t))^T \mathbf{R}_t^{-1}(\mathbf{y}_t - H_t(\mathbf{x}_t))\right], \tag{15}$$

$$P(\mathbf{x}_t|\mathbf{x}_{t-1}) = (2\pi)^{-\frac{n}{2}}|\mathbf{Q}_t|^{-\frac{1}{2}} exp\left[-\frac{1}{2}(\mathbf{x}_t - M(\mathbf{x}_{t-1}))^T \mathbf{Q}_t^{-1}(\mathbf{x}_t - M(\mathbf{x}_{t-1}))\right]. \tag{16}$$

By substituting equations (14)–(16) into equation (13), ignoring constant terms, and combining the negative sign from equation (13) with the signs from the Normal distributions, the cost function for DA can be expressed as:

$$\mathbf{x} = \underset{\mathbf{x}}{\operatorname{argmin}} \frac{1}{2}\left[(\mathbf{x}_0 - \mathbf{x}_b)^T \mathbf{B}^{-1}(\mathbf{x}_0 - \mathbf{x}_b) + \sum_{i=1}^{T}(\mathbf{y}_t - H_t(\mathbf{x}_t))^T \mathbf{R}_t^{-1}(\mathbf{y}_t - H_t(\mathbf{x}_t)) + \sum_{t=1}^{T}(\mathbf{x}_t - M(\mathbf{x}_{t-1}))^T \mathbf{Q}_t^{-1}(\mathbf{x}_t - M(\mathbf{x}_{t-1}))\right]. \tag{17}$$

This can be represented as a cost function using weighted L2 norms:

$$J(\mathbf{x}) = \| \mathbf{x}_0 - \mathbf{x}_b \|^2_{\mathbf{B}^{-1}} + \sum_{t=1}^{T} \| \mathbf{y}_t - H(\mathbf{x}_t) \|^2_{\mathbf{R}_t^{-1}} + \sum_{t=1}^{T} \| \mathbf{x}_t - M(\mathbf{x}_{t-1}) \|^2_{\mathbf{Q}_t^{-1}}. \tag{18}$$

Analytical solutions to equations (12) and (13) can be derived only for a few distributions, such as Normal and Laplace distributions. For more general distributions, the Markov chain Monte Carlo (MCMC) method is commonly employed to generate a large number of sample sets to approximate the probability distributions. Additionally, variational methods such as four-dimensional variational methods are often used to solve equations (17) and (18), which require explicit computations of adjoint model and Hessian matrix of the dynamical model and observation model.

2.3.2. Loss/cost function of ML

Similarly, the MAP estimation method can be used to estimate the parameter in equation (11). In some literature, such as Xu and Sun (2018), this is also expressed as an estimation of the parametric model $F_{\boldsymbol{\omega}}$, both of which are equivalent. Thus, the argmax of equation (11) is then expressed as:

$$\boldsymbol{\omega} = \underset{\boldsymbol{\omega}}{\operatorname{argmax}}\, P(\boldsymbol{\omega}|\mathcal{D}) = \underset{\boldsymbol{\omega}}{\operatorname{argmax}}\, P(\mathbf{x}|\mathbf{y}, \boldsymbol{\omega})P(\boldsymbol{\omega}) = \underset{\boldsymbol{\omega}}{\operatorname{argmax}} \prod_{t=0}^{T} P(\mathbf{x}_t|\mathbf{y}_t, \boldsymbol{\omega})P(\boldsymbol{\omega}). \tag{19}$$

By taking the logarithm of both sides of the equation and adding a negative sign, we convert the maximization problem into a minimization problem.

If the likelihood $P(\mathbf{x}_t|\mathbf{y}_t, \boldsymbol{\omega})$, and $P(\boldsymbol{\omega})$ follow multivariate Normal distributions, i.e., $\mathbf{x}_t|\mathbf{y}_t \sim \mathcal{N}(F_{\boldsymbol{\omega}}(\mathbf{y}_t), \mathbf{I})$, and $\boldsymbol{\omega} \sim \mathcal{N}(\tilde{\boldsymbol{\omega}}, \boldsymbol{\sigma}_{\boldsymbol{\omega}})$, where, the covariance matrix of $P(\mathbf{x}_t|\mathbf{y}_t, \boldsymbol{\omega})$ is assumed to be the identity matrix, and $\tilde{\boldsymbol{\omega}}$ and $\boldsymbol{\sigma}_{\boldsymbol{\omega}}$ are the mean and covariance matrix of the parameter set. Both covariance matrices are not time-dependent, meaning that in ML, estimates are made based on all time as a whole. The specific forms of their multivariate Normal distributions are not elaborated here.

Thus, we have

$$\underset{\boldsymbol{\omega}}{\operatorname{argmin}} \frac{1}{2}\left[\sum_{t=0}^{T}(\mathbf{x}_t - F_{\boldsymbol{\omega}}(\mathbf{y}_t))^T(\mathbf{x}_t - F_{\boldsymbol{\omega}}(\mathbf{y}_t)) + \sum_{t=0}^{T}(\boldsymbol{\omega} - \tilde{\boldsymbol{\omega}})^T \boldsymbol{\sigma}_{\boldsymbol{\omega}}^{-1}(\boldsymbol{\omega} - \tilde{\boldsymbol{\omega}})\right]. \tag{20}$$

Note that $\mathbf{y}$ represents the input observational data, while $\mathbf{x}$ represents the targeted/labeled output data. The above equation can also be represented as a loss/cost function using weighted L2 norms:

$$L(\boldsymbol{\omega}) = \sum_{t=0}^{T} \| \mathbf{x}_t - F_{\boldsymbol{\omega}}(\mathbf{y}_t) \|^2 + \sum_{t=0}^{T} \| \boldsymbol{\omega} - \tilde{\boldsymbol{\omega}} \|^2_{\boldsymbol{\sigma}_{\boldsymbol{\omega}}^{-1}}. \tag{21}$$

The last term in the equation is a regularization term, which helps prevent overfitting by incorporating prior information about the parameter values.

When the likelihood and prior are not Gaussian, the analytical solution for $\boldsymbol{\omega}$ may not be straightforward. In such cases, numerical optimization methods like backward propagation and stochastic gradient descent can be employed to find the parameter estimates. Additionally, sampling-based approaches such as MCMC techniques can be used to approximate the posterior distribution. These methods do not rely on the Gaussian assumption and are capable of handling complex, nonlinear, and non-Gaussian relationships in the data, thus providing a flexible framework for posterior inference.

Comparing equations (17) and (20) (or equations (18) and (21)) reveals a similarity in the cost function structures between DA and ML. The main difference is that DA's cost function explicitly includes the dynamical model M and the observation model H in cost function of DA. ML's loss function does not include system dynamics and observation model explicitly. However, this does not mean that ML algorithms cannot incorporate dynamics. For example, physics-informed ML proposes regularizations (Karniadakis et al., 2021) to introduce constraints or penalties within the cost function, and then capture the underlying patterns and prevent poor generalization and overfitting. Additionally, the relationships between system states and observations in DA and ML are opposite, in DA, system states are transferred to observation space using an observation model, whereas in ML, observational data are transferred to system states using a parametric model.

3. Distinctions between DA and ML in earth system states estimation and prediction

DA and ML utilize similar cost function frameworks, as discussed mathematically at the above section, yet they fundamentally differ in

Table 1
Comparative analysis of data assimilation (DA) and machine learning (ML) in Earth system state estimation and prediction.

	DA	ML
Motivation and application scope	To improve state estimation and prediction accuracy by fusing observational data into physical-based models.	To improve prediction by estimating the model parameter from large datasets, with a lesser emphasis on dynamical systems.
Governing equations of dynamics and observations	Known system dynamics and observation model $\mathbf{x}_t = M(\mathbf{x}_{t-1}) + \eta_t$ $\mathbf{y}_t = H(\mathbf{x}_t) + \boldsymbol{\varepsilon}_t$	No explicit system dynamics and observation model
Parameters	Parameters are usually with physical meanings and interpretable.	Parameters are related to model structure and often abstracted from data, with varied interpretability.
Cost functions	$\mathbf{x} = \underset{\mathbf{x}}{\operatorname{argmax}}\, P(\mathbf{x}|\mathbf{y}_{1:T})$ $J(\mathbf{x}) = \| \mathbf{x}_0 - \mathbf{x}_b \|^2_{\mathbf{B}^{-1}} + \sum_{t=1}^{T} \| \mathbf{y}_t - H(\mathbf{x}_t) \|^2_{\mathbf{R}_t^{-1}} + \sum_{t=1}^{T} \| \mathbf{x}_t - M(\mathbf{x}_{t-1}) \|^2_{\mathbf{Q}_t^{-1}}$	$\boldsymbol{\omega} = \underset{\boldsymbol{\omega}}{\operatorname{argmax}}\, P(\boldsymbol{\omega}|\mathcal{D})$ $L(\boldsymbol{\omega}) = \sum_{t=0}^{T} \| \mathbf{x}_t - F_{\boldsymbol{\omega}}(\mathbf{y}_t) \|^2 + \sum_{t=0}^{T} \| \boldsymbol{\omega} - \tilde{\boldsymbol{\omega}} \|^2_{\boldsymbol{\sigma}_{\boldsymbol{\omega}}^{-1}}$
Optimization algorithms	Uses Bayesian filtering and variational methods to integrate data into models, often requiring explicit computations of adjoint model and Hessian matrix.	Employs algorithms like backward propagation and stochastic gradient descent in optimization.
Uncertainty quantification	Explicitly formulated and quantified in the DA process.	Embedded within the model structure, and implicitly handled.
Data utilization	Requires sparse observational data with high fidelity.	Relies on the availability of large-scale dense data.
Computational characteristics	CPU-based parallel computing for ensemble processing; scales with ensemble size.	GPU-accelerated computing for network training; scales with data volume.

underlying principles and applications in Earth system state estimation and prediction. These differences are outlined in Table 1 and elucidated in the following discussion.

Motivation and application scope. DA focuses on more accurate state estimations by actively incorporating current and past observational data into physical-based models. DA excels within domains grounded in rigorous physical theories, such as atmospheric, oceanic, and terrestrial systems dynamics. On the other hand, ML's predictive ability stems from its capacity to discern complex patterns and relationships in large amounts of data. While ML models sometimes lack a clear physical basis, they excel in identifying and leveraging correlations within big data to make predictions.

Governing equations and parameters. The integration of dynamical and observation models within DA relies on established physical equations, usually differential equations, facilitating detailed representations of spatiotemporal evolutions and the relationships between state variables and observations. Additionally, parameters in dynamical and observation models have physical meanings and can be derived from or constrained by physical measurements and theoretical models. ML, on the other hand, often operates devoid of explicit physics-based governing equations, confining its efficacy to situations where data alone can reveal trends and patterns that are used for making estimations and predictions, even though this may sometimes lead to reduced interpretability of how these estimations and predictions are made.

Cost functions. DA's cost function incorporates both dynamical and observation models, ensuring that state estimations adhere to physical laws and observed data. Conversely, ML focuses on optimizing prediction accuracy through data-driven methods without explicitly embedding such models. In combining system states with observations, DA transforms system states into the observation space, whereas ML uses a parametric model to map observational data back to system states. These differences underscore the complementary strengths of DA's adherence to physical principles and ML's flexibility, pointing to potential integration for enhancing predictive capabilities.

Optimization algorithms. DA places emphasis on the integration of observational data with model dynamics. This integration is often facilitated by techniques including variational methods such as 3-dimensional and 4-dimensional variational algorithms, along with Bayesian filtering methods like the Kalman filter, ensemble Kalman filter, and particle filter (Evensen, 2009), hybrid approaches like Ensemble 4-Dimensional Variational (En4DVar) methods also play a significant role (Liu et al., 2009). These methods often require explicit computations of adjoint model and Hessian matrix or apply MCMC algorithms. On the counterpart, ML refines the architecture of models and fine-tunes parameters for predictions. Optimization algorithms such as backward propagation and stochastic gradient descent are central to ML's approach, enabling intricate model adjustments. These methods offer insights for the relatively less efficient computations of the adjoint model and Hessian matrix in DA. Particularly in addressing nonlinear or non-Gaussian problems, adopting backward propagation and stochastic gradient descent algorithms could enhance the computational efficiency of state updating and parameter estimation within DA.

Uncertainty quantification. In DA, model uncertainties, observation uncertainties, and analysis uncertainties are explicitly formulated and quantified in the assimilation process through covariance matrices that weigh the influence of observations and model predictions. The Kalman filter, for example, strikes an optimal balance between model and observational errors to improve prediction accuracy, while variational methods incorporate these uncertainties directly in their cost functions. In ML, uncertainties are embedded within the model structure and handled implicitly through weight adjustments. This is achieved through regularization techniques in the cost function to prevent overfitting, and through Bayesian approaches that learn the posterior distribution of parameters. These mechanisms in both DA and ML ensure that model parameters are optimized to enhance accuracy and generalizability while accounting for uncertainties.

Data utilization. DA uses sparse but usually high-fidelity observational data. Subsequently, DA generates physically consistent and spatiotemporally continuous data by combing the strength of physical-based dynamics with sparse data. On the contrary, ML depends on extensive training and learning from large-scale dense data to make predictions. However, advancements in Generative Adversarial Networks (GANs) and reinforcement learning are expanding ML's capability to efficiently utilize sparse data, which is critically important for Earth system predictions inherently limited by data sparsity.

Computational characteristics. From a computational perspective, DA and ML diverge substantially in their hardware and software ecosystems. DA systems, often built on Fortran or Python/C++ frameworks like DART and PDAF, rely on HPC clusters to manage ensemble-based uncertainty quantification and adjoint models. These workflows prioritize low-latency, high-throughput parallelism across CPU nodes, balancing ensemble size with spatial resolution. In contrast, ML leverages GPU-accelerated Python frameworks (e.g., PyTorch) for training deep networks on massive datasets, where computational efficiency is driven by matrix optimizations and distributed data parallelism. While DA's costs scale with model complexity and ensemble size, ML's scalability hinges on data volume and network depth. Bridging a new paradigm—such as integrating ML emulators into DA workflows or using DA to generate training data—requires hybrid architectures that harmonize CPU-GPU resources and algorithmic interoperability.

4. Discussion and conclusions

This work examines DA and ML in Earth system states estimation and prediction through a mathematical lens grounded in Bayesian inference, elucidating their objective coherence and cost function derivation. Despite their shared similarities in Bayesian foundations, their implementation varies distinctly across aspects like motivations, governing equations, parameterization, cost functions, algorithms, uncertainty handling, data demands, and computational characteristics. DA and ML offer complementary benefits in Earth system state estimation and prediction. DA excels in interpretability stemming from physical principles, whereas ML offers enhanced operationality through its data-centric approach. Consequently, the integration of DA and ML will likely autonomously contribute to the discovery of new scientific knowledge (Li and Guo, 2025).

DA can improve its operationality by borrowing ML's strong ability in addressing the nonlinearity and non-Gaussian challenges (Brajard et al., 2021; Zhao et al., 2019), for example, adopting backward propagation and stochastic gradient descent in tackling the nonlinearity problem and difficulties in developing adjoint models in four-dimensional variational DA. Additionally, DA can substantially enhance its efficiency by introducing ML to develop efficient surrogate model or DA system, therefore representing complex physical processes without intensive computational demands. For example, latent data assimilation performs a reduced-order surrogate model with high algorithmic efficiency (Amendola et al., 2021), and meanwhile revealing unobserved dynamics, and thus enhancing model fidelity.

ML can greatly benefit from DA's rigorous theoretical frameworks. By integrating DA's dynamical models and theoretical insights, ML can achieve a more balanced theory-data approach. This integration helps ML enhance its interpretability and build more robust predictive models. DA not only provides critical prior knowledge but also generates detailed reanalysis datasets, enriching the diversity and depth of training data for ML applications. Advanced models like FourCastNet and Pangu-Weather demonstrate the power of utilizing comprehensive reanalysis datasets such as ERA5 (Hersbach et al., 2020), substantially improving prediction accuracy and capability (Bi et al., 2023; Pathak et al., 2022). This dependence of current ML-based models on DA highlights the importance of DA systems as key sources of training data. The performance of ML models relies heavily on the quality and availability of DA outputs, which are essential for learning patterns and relationships in the data. This underscores the need for hybrid approaches that combine ML's strengths in handling large

datasets and identifying complex patterns with the stability and reliability of DA systems for enhanced Earth system state estimation.

Moreover, ML can adopt DA's emphasis on understanding and quantifying uncertainty within both models and data (Buizza et al., 2022). ML can adopt DA's rigorous approach to address uncertainties, an aspect often not fully considered in standard ML practices. Additionally, ML can learn effective strategies from DA to address the challenge of overfitting, a common pitfall in model training. While traditional ML solutions like bias-variance tradeoff and regularization address overfitting, DA leverages prior knowledge to augment data, enhancing the model's ability to focus on meaningful patterns and reduce the impact of random errors and noise. This allows ML models to refine their focus on relevant information, leading to more accurate and generalizable predictions.

The integration of DA and ML presents unique opportunities in addressing the challenge of data sparsity that is inherent in Earth system science. Specifically, the advancements in GANs and reinforcement learning have widened the scope of ML, enabling a more efficient utilization of sparse data. Therefore, integrating these cutting-edge ML techniques into DA frameworks could lead to breakthroughs in Earth system prediction, even in scenarios where observational data are scarce.

The integration of DA and ML also offers a transformative pathway to address Earth system modeling challenges (Pan et al., 2025). By merging DA's physics-based uncertainty quantification with ML's data-driven pattern recognition, hybrid frameworks can achieve both interpretability and scalability. For instance, ML emulators trained on DA outputs (e.g., ERA5 reanalysis) have demonstrated success in accelerating high-resolution weather forecasting (Bi et al., 2023), while DA systems enhanced by ML-generated covariance matrices improve state estimation efficiency (Brajard et al., 2021). Key challenges remain in harmonizing their computational ecosystems—bridging DA's HPC-dependent ensembles with ML's GPU-optimized training—and ensuring theoretical consistency when coupling dynamical models with ML components. Addressing these challenges will pave the way for robust digital twins capable of real-time Earth system monitoring and prediction (Li et al., 2023).

In conclusion, to advance Earth system science states estimation and prediction, it is critical to develop an integrated approach that combines the theoretical rigor of DA with the ability of ML to leverage vast data. However, challenges persist. The ongoing discourse surrounds the optimal framework that unites DA's theoretical strengths with ML's computational prowess.

CRediT authorship contribution statement

Xin Li: Writing – review & editing, Writing – original draft, Visualization, Supervision, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Feng Liu:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Funding acquisition, Formal analysis.

Funding

This work is supported by the National Natural Science Foundation of China (grant numbers 42430112 & 42271442).

Declaration of competing interest

The authors declare no conflict of interest.

References

Amendola, M., Arcucci, R., Mottet, L., Casas, C.Q., Fan, S., Pain, C., et al., 2021. Data Assimilation in the Latent Space of a Convolutional Autoencoder. Springer International Publishing.

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., Tian, Q., 2023. Accurate medium-range global weather forecasting with 3D neural networ ks. Nature. https://doi.org/10.1038/s41586-023-06185-3.

Brajard, J., Carrassi, A., Bocquet, M., Bertino, L., 2021. Combining data assimilation and machine learning to infer unresolved s cale parametrization. Phil. Trans. Math. Phys. Eng. Sci. https://doi.org/10.1098/rsta.2020.0086.

Buizza, C., Quilodrán Casas, C., Nadler, P., Mack, J., Marrone, S., Titus, Z., et al., 2022. Data learning: integrating data assimilation and machine learning. J. Comput. Sci. https://doi.org/10.1016/j.jocs.2021.101525.

Carrassi, A., Bocquet, M., Bertino, L., Evensen, G., 2018. Data assimilation in the geosciences: an overview of methods, issues, and perspectives. WIREs Clim. Change. https://doi.org/10.1002/wcc.535.

Course, K., Nair, P.B., 2023. State estimation of a physical system with unknown governing equations. Nature. https://doi.org/10.1038/s41586-023-06574-8.

Evensen, G., 2009. The ensemble Kalman filter for combined state and parameter estimation. IEEE Control Syst. https://doi.org/10.1109/MCS.2009.932223.

Gao, T.-T., Barzel, B., Yan, G., 2024. Learning interpretable dynamics of stochastic complex systems from experimental data. Nat. Commun. https://doi.org/10.1038/s41467-024-50378-x.

Geer, A.J., 2021. Learning earth system models from observations: machine learning or da ta assimilation? Phil. Trans. Math. Phys. Eng. Sci. https://doi.org/10.1098/rsta.2020.0089.

Gettelman, A., Geer, A.J., Forbes, R.M., Carmichael, G.R., Feingold, G., Posselt, D.J., et al., 2022. The future of Earth system prediction: advances in model-data fusion. Sci. Adv. https://doi.org/10.1126/sciadv.abn3488.

Ghahramani, Z., 2015. Probabilistic machine learning and artificial intelligence. Nature. https://doi.org/10.1038/nature14541.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. The MIT Press.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., et al., 2020. The ERA5 global reanalysis. Q. J. R. Meteorol. Soc. https://doi.org/10.1002/qj.3803.

Karniadakis, G.E., Kevrekidis, I.G., Lu, L., Perdikaris, P., Wang, S., Yang, L., 2021. Physics-informed machine learning. Nat. Rev. Phys. https://doi.org/10.1038/s42254-021-00314-5.

Kochkov, D., Yuval, J., Langmore, I., Norgaard, P., Smith, J., Mooers, G., et al., 2024. Neural general circulation models for weather and climate. Nature. https://doi.org/10.1038/s41586-024-07744-y.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., et al., 2023. Learning skillful medium-range global weather forecasting. Science. https://doi.org/10.1126/science.adi2336.

Li, X., Guo, Y.L., 2025. Paradigm shifts from data-intensive science to robot scientists. Sci. Bull. https://doi.org/10.1016/j.scib.2024.09.029.

Li, X., Liu, F., Fang, M., 2020. Harmonizing models and observations: data assimilation in Earth system science. Sci. China Earth Sci. https://doi.org/10.1007/s11430-019-9620-x.

Li, X., Feng, M., Ran, Y., Su, Y., Liu, F., Huang, C., et al., 2023. Big data in Earth system science and progress towards a digital twin. Nat. Rev. Earth Environ. https://doi.org/10.1038/s43017-023-00409-w.

Li, X., Liu, F., Ma, C., Hou, J., Zheng, D., Ma, H., et al., 2024. Land data assimilation: harmonizing theory and data in land surface process studies. Rev. Geophys. https://doi.org/10.1029/2022RG000801.

Liu, C., Xiao, Q., Wang, B., 2009. An ensemble-based four-dimensional variational data assimilation scheme. Part II: observing system simulation experiments with advanced rese arch WRF (ARW). Mon. Weather Rev. https://doi.org/10.1175/2008MWR2699.1.

McLaughlin, D., 1995. Recent developments in hydrologic data assimilation. Rev. Geophys. https://doi.org/10.1029/95RG00740.

Pan, X.D., Chen, D.L., Pan, B.X., Huang, X.Z., Yang, K., Piao, S.L., Zhou, T.J., Dai, Y.J., Chen, F.H., Li, X., 2025. Evolution and prospects of Earth system models: challenges and opportunities. Earth Sci. Rev. https://doi.org/10.1016/j.earscirev.2024.104986.

Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., et al., 2022. FourCastNet: a global data-driven high-resolution weather model using adaptive fourier neural operators. https://doi.org/10.48550/arXiv.2012.12056.

Penny, S.G., Smith, T.A., Chen, T.C., Platt, J.A., Lin, H.Y., Goodliff, M., et al., 2022. Integrating recurrent neural networks with data assimilation for scalable data-driven state estimation. J. Adv. Model. Earth Syst. https://doi.org/10.1029/2021MS002843.

Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., et al., 2021. Skilful precipitation nowcasting using deep generative models of radar. Nature. https://doi.org/10.1038/s41586-021-03854-z.

Wikle, C.K., Berliner, L.M., 2007. A Bayesian tutorial for data assimilation. Phys. Nonlinear Phenom. https://doi.org/10.1016/j.physd.2006.09.017.

Xu, Z., Sun, J., 2018. Model-driven deep-learning. Natl. Sci. Rev. https://doi.org/10.1093/nsr/nwx099.

Zhang, Y., Long, M., Chen, K., Xing, L., Jin, R., Jordan, M.I., et al., 2023. Skilful nowcasting of extreme precipitation with NowcastNet. Nature. https://doi.org/10.1038/s41586-023-06184-4.

Zhao, W.L., Gentine, P., Reichstein, M., Zhang, Y., Zhou, S., Wen, Y., et al., 2019. Physics-constrained machine learning of evapotranspiration. Geophys. Res. Lett. https://doi.org/10.1029/2019GL085291.