Learning-based inversion-free model-data integration to advance ecosystem model prediction

Dan Lu
Oak Ridge National Laboratory
Oak Ridge, TN, USA
lud1@ornl.gov

Daniel Ricciuto
Oak Ridge National Laboratory
Oak Ridge, TN, USA
ricciutodm@ornl.gov

***Abstract*—Ecosystem model prediction is important for understanding ecosystem responses to climate change and for management support. Model prediction and quantification of predictive uncertainty of ecosystems have long been investigated. The traditional workflow, which calibrates models to match observations and then uses the calibrated models for predictions, relies heavily on inverse modeling to constrain uncertain parameters in complex forward models. This inversion-based prediction approach is infeasible for complex models with heterogeneous parameter uncertainties and incapable of rapid integration of streaming and multiple sources of data because of the difficulty and computational cost in the model inversion, which is typically ill-posed and can require hundreds of thousands of expensive forward simulations to be performed iteratively. We propose to circumvent inverse modeling by precomputing an ensemble of unconstrained forward simulations and then using machine learning (ML) methods to learn the statistical relationship between simulated observation and prediction quantities. Once the ML model has learned the relationship, it can be used to make predictions of future system behavior with uncertainty quantification based on observations. The proposed learning-based inversion-free model prediction (LIMP) framework is computationally efficient which only requires a few thousands of fully parallelizable forward simulations. Additionally, LIMP can continually update predictions based on streaming observations from multiple locations and sources without necessarily requiring extra model simulations. In this study, we apply LIMP to a regional terrestrial ecosystem model with 47 parameters for testing, refining, and evaluating the approach. We demonstrate that LIMP can be used for efficient model prediction, rapid data assimilation, and cost-effective experimental design for improving robust predictive understanding of ecosystems.**

***Index Terms*—Efficient model prediction, Rapid data assimilation, Machine learning, Uncertainty quantification, Terrestrial ecosystem modeling**

I. Introduction

In the face of unprecedented global climate change, there is growing demand for the prediction of ecosystem responses that provides actionable information for policy making and management support. Currently, uncertainties associated with ecosystem responses are so great that it is unknown, given a scenario describing future anthropogenic emissions of carbon dioxide (CO_2) to the atmosphere, whether the terrestrial biosphere will be a sink or a source of atmospheric CO_2 in the latter half of the 21st century. Answering this question requires accurate prediction of net CO_2 exchange of terrestrial ecosystems.

Traditional model prediction methods utilize inference on model parameters from observation data to make predictions. These traditional approaches typically involve an iterative model inversion by tuning model parameters to fit the observations and then use the calibrated model to predict quantities of interest (QoIs). While several studies have shown that these inversion-based methods improved the terrestrial ecosystem predictions [1], [2], application of these methods to complex models has several computational challenges. First, a complex model typically involves a large number of model parameters exhibiting heterogeneous uncertainties, e.g., categorical uncertainty about which process model for use, spatial uncertainty of gridded model variables, and continuous uncertainty of continuous parameters. No model inversion methods can accurately estimate all these diverse parameters and quantify their heterogeneous uncertainties. Second, the complex and nonlinear relationship between the model parameters and observation variables makes the iterative model inversion time-consuming and computation-demanding. For example, a global optimization method such as genetic algorithms and Markov chain Monte Carlo sampling can easily require hundreds of thousands of iterative model evaluations, and the number of model runs increases exponentially with the parameter size and the model nonlinearity. Many applications cannot afford such a large amount of computing costs and default parameter values are usually used, which impairs the prediction performance. Third, the dimension of model parameters can be larger than that of observation variables, making the inverse problem under-determined and ill-posed. Last but not the least, these inversion-based computationally expensive model prediction procedure needs to be repeated to integrate new observations or whenever the objective function changes, which makes it infeasible for rapid data assimilation and incapable of robust model-data integration.

Surrogate modeling methods have been developed to address these computational challenges [3], [4]. By building a cheap-to-evaluate surrogate of the original computationally expensive model and replacing the original model with the surrogate model in the traditional approach, we can reduce the computing time in the iterative model calibration and succeeding model prediction. However, building an accurate surrogate itself involves a large amount of computation. For example, using a pth-order polynomial to approximate a model

with d parameters, $M = (p+d)!/(p!d!)$ coefficients need to be solved, i.e., the number of coefficients increases factorially fast with the parameter size and the polynomial order. A terrestrial ecosystem model typically has dozens of uncertain parameters and a high-order polynomial is usually needed for approximation accuracy. This can easily lead to a prohibitive number of model evaluations, up to 10^5, necessary for surrogate construction. More importantly, surrogate modeling does not solve the fundamental problems of the traditional prediction methods, such as ill-posedness, heterogeneous parameter uncertainties and iterative inversion, attributed to the methods' nature in making predictions from parameter inference.

In this work, we propose a learning-based inversion-free model prediction (LIMP) framework that can efficiently produce accurate predictions and predictive uncertainty informed from observation data through forward simulations. The key idea of LIMP is to circumvent the challenging inverse modeling by precomputing an ensemble of unconstrained forward simulations and then using machine learning (ML) methods to learn the statistical relationship between simulated observations and predictive QoIs. Once the ML model has learned the relationship, it can be used to continually update predictions of future system behavior based on streaming and multiple sources of observation data to enable rapid data assimilation. Specifically, we first use Monte Carlo sampling to simulate heterogeneous uncertainties of model parameters based on their prior distributions, and then evaluate forward models to generate corresponding prior realizations of observation variables (e.g., historical carbon flux measurements at some sites) and prediction variables (e.g., future carbon fluxes in a region). Next, we use ML models to learn the observation-prediction relationship based on their realizations. Dimension reduction techniques are used to reduce the dimensionality of simulated observations and predictions to facilitate the construction of their relationship. Last, we use Bayesian inference to infer the prediction probability distribution according to the established relationship with actual field observation data.

Compared to the traditional model prediction methods, LIMP has the following merits. First, LIMP considers model structural and parameter uncertainties through forward simulations in a non-iterative and fully-parallel way, which enables LIMP to consider heterogeneous uncertainties in a computationally efficient manner. Second, LIMP infers predictions from established statistical relationships between observations and predictions without explicit model inversion, which avoids the aforementioned difficulties in the inverse modeling. Third, LIMP works on the observation-prediction space whose relationship is relatively simple and whose dimensionality can be greatly reduced because most observation and prediction variables in Earth system science are time-series or spatial-map responses having strong spatiotemporal correlations. This enables LIMP to easily establish the observation-prediction relationship using a wide range of dimension reduction and statistical regression techniques. Fourth, LIMP uses Bayesian inference to infer the prediction which incorporates prior information and produces sophisticated uncertainty quantification. However, different from the traditional Bayesian inference where the likelihood is formulated on the model parameters and computed on the actual physical models, the likelihood in LIMP is defined on the prediction variables in a reduced dimension and computed on a statistical model. This definition of the likelihood allows LIMP to infer the Bayesian prediction directly and efficiently without the traditionally required computationally-intensive Markov chain Monte Carlo sampling. Last, LIMP is able to make continuous updating of predictions by integrating potentially near real-time new observation data without necessarily requiring extra model simulations, which significantly improves the computational efficiency and enables the fast model-data integration.

In this study, we apply the LIMP framework to a regional terrestrial ecosystem model with 47 parameters for testing, refining, and evaluating the approach, and demonstrate that the proposed method can be used for efficient model prediction, rapid data assimilation and cost-effective experimental design for improving robust predictive understanding of ecosystems. In the following, we first introduce the LIMP framework. Next, we describe the terrestrial ecosystem model and the model prediction and data assimilation problems. We then apply LIMP to solve the problems and analyze the results. Last, we discuss limitations of the proposed method and provide directions for future development.

II. Methodology

In this section, we introduce the LIMP framework. We first describe the general procedure of LIMP and then explain its specific components in detail.

A. General procedure of the method

The key of LIMP is to establish an observation-prediction relationship from their prior samples in a reduced dimension to be able to estimate the posterior prediction distribution for given observation data. Specifically, LIMP consists of four steps:

1) Generating prior samples of observation and prediction variables by running forward models based on the prior distribution of model parameters;
2) Dimension reduction of the simulated observations and predictions;
3) Establishing a statistical relationship between observation and prediction in a reduced dimension;
4) Bayesian inference of the prediction based on the statistical model with observation data.

Steps 1-3 correspond to the training stage, where the observation-prediction relationship in the reduced dimension is learned from unconstrained forward simulations. Step 4 corresponds to the prediction stage, where the posterior distribution of the prediction is deduced from the observed data after back transformation to its original high-dimensional space.

B. Generation of prior samples

The first step of establishing a statistical relationship between observation and prediction is the collection of samples

of both variables in a systematic way such that they can be used to train or fit the statistical model. Here we denote the observation variables as $\mathbf{d}$ and the prediction variables as $\mathbf{h}$. Both $\mathbf{d}$ and $\mathbf{h}$ are system responses, they can be a time-series response at t different time-steps or a spatial map at ℓ different locations or variables varying in both spatial and temporal spaces. Generally, $\mathbf{d}$ and $\mathbf{h}$ have high dimensionality and can be simulated by a physics-based model (such as terrestrial ecosystem models in this study). The model simulating $\mathbf{d}$ and the model simulating $\mathbf{h}$ can be the same or can be different (e.g., under different future CO_2 emission scenarios). Both models have uncertain parameters whose uncertainty can be quantified using some prior distributions constrained in a physically-reasonable range. The prior parameter distributions can be uniform, Gaussian or any other distributions dependent on our prior knowledge about the parameters. After stating prior distributions, a set of N samples of parameters are generated. Through forward model simulations, we obtain sets of samples of $\mathbf{d}$ and $\mathbf{h}$. Note that the samples of $\mathbf{d}$ and $\mathbf{h}$ can be from one model or multiple models. Considering multiple model samples quantifies model structural uncertainty and helps improve predictive performance. Ideally, the model simulating $\mathbf{d}$ and the model simulating $\mathbf{h}$ should share the same uncertain parameters to facilitate establishing the observation-prediction relationship and constraining the prediction from the observation data. If the prediction model has more uncertain parameters than the observation model, even though we can build a nice statistical relationship from their prior samples, the prediction may not be constrained well by the observation data, resulting in inaccurate predictions with large uncertainty. Specifying appropriate priors is important for Bayesian inference but also challenging and still an active field of research. In section IV, we will discuss the impact of priors on the LIMP performance.

C. Dimension reduction

Response variables $\mathbf{d}$ and $\mathbf{h}$ usually have spatial or/and temporal correlations. When the variable dimensions are highly correlated with each other, multicollinearity occurs. Multicollinearity results in numerical issues during model fitting and degrades the predictive performance of the statistical model. Dimension reduction identifies the degrees of freedom that capture the majority of the variance in the data. Therefore, performing statistical analysis in the reduced dimension removes the multicollinearity and facilitates the model fitting. Additionally, dimension reduction reduces the variables and thus reduces the required number of samples, which improves the computational efficiency and enhances the model reliability.

In this study, we use principal component analysis (PCA) for dimension reduction. PCA is a multivariate analysis technique that applies an orthogonal transformation to convert a set of samples of possibly correlated variables into a set of values of uncorrelated variables, called principal components. Typically, the first K components of the PCA decomposition, with $K \ll N$, explain almost all the variance of the data. By keeping only those K first dimensions, we thus can achieve an efficient dimension reduction. PCA identifies the principal modes of variation from the eigen-vectors of the covariance matrix. When the observation and prediction variables represent systematic signals varying in time, we can use functional PCA (FPCA) for dimension reduction. FPCA first uses functional data analysis to decompose the time series data into a linear combination of basis functions such as polynomial basis and spline basis, and then uses PCA on the coefficients of this linear combination to characterize the functional variations in the time series data. For the prediction variable $\mathbf{h}(t)$, the FPCA can be written as:

$$\mathbf{h}(t) \cong \sum_{k=1}^{K} h_k^f \phi_{h,k}(t), \tag{1}$$

where $\boldsymbol{\phi}_h(t)$ represents the K orthonormal eigen-functions; $\mathbf{h}^f$ represents the first K principal components, i.e., the $\mathbf{h}$ variables in the reduced dimension. A similar decomposition can be achieved for the observation variables as,

$$\mathbf{d}(t) \cong \sum_{k=1}^{K} d_k^f \phi_{d,k}(t). \tag{2}$$

When the observation and prediction variables are spatial response maps, we can apply eigen-image analysis for dimension reduction in the similar manner of Eq. 1 and 2, which is a spatial version of the FPCA. When the observation data are from multiple sources, e.g., from different locations where each location data is a time series, or from different types where each type of data is a time series/spatial map, we can use a mixed PCA to pool data together and generate a reduced dimensional projection of the combined data. First, a FPCA is performed on each of the data sources to obtain the largest singular values. Next, each data source is normalized according to its first singular value; this accounts for any difference in scales amongst the data sources. Last, the normalized data inputs are concatenated and the standard PCA is applied to this final matrix. After applying dimension reduction to the set of N prior samples of the observation and prediction, we obtain N samples of $\mathbf{d}^f$ in the reduced dimension p, and N samples of $\mathbf{h}^f$ in the reduced dimension q, where p and q can be different according to the functional properties of $\mathbf{d}$ and $\mathbf{h}$.

Both PCA, FPCA, and mixed PCA are bijective operations, meaning that the original high-dimensional variable can be recovered by undoing the projection. This allows for first establishing statistical models in the low-dimensional space easily and efficiently, and then any predictions made in the low dimension can be reconstructed in the original space uniquely. The precision of the reconstruction depends on the number of components used. For cases where the majority of the variance of the data is captured by the first few components, accurate reconstruction can be expected using only those first few component bases with a small loss of information.

D. Establishing the statistical relationship

The relationship between $\mathbf{d}^f$ and $\mathbf{h}^f$ in the reduced dimension can be nonlinear which challenges the statistical model

learning. We can first use canonical correlation analysis (CCA) to linearize the relationship to simplify the model fitting. CCA is a multivariate analysis method that can be applied to transform the relationships between pairs of vector variables with N samples into a set of independent linearized relationships between pairs of scalar variables [5]. For example, we have two vectors $\mathbf{d}^f = (\mathbf{d}_1^f, \ldots, \mathbf{d}_p^f)$ and $\mathbf{h}^f = (\mathbf{h}_1^f, \ldots, \mathbf{h}_q^f)$ of random variables with N samples, and $\mathbf{d}^f$ and $\mathbf{h}^f$ are correlated. CCA finds linear combinations of $\mathbf{d}^f$ and $\mathbf{h}^f$ which have maximum correlation with each other. Mathematically, CCA seeks vectors $\mathbf{a}_i(\mathbf{a}_i \in \mathbb{R}^p)$ and $\mathbf{b}_i(\mathbf{b}_i \in \mathbb{R}^q)$ with $i = 1 \cdots m$ that maximize the correlation between the linear combinations $\mathbf{a}_i^T \mathbf{d}^f$ and $\mathbf{b}_i^T \mathbf{h}^f$. The resulting linear combinations are denoted as $\mathbf{d}_i^c$ and $\mathbf{h}_i^c$, and called the canonical variates of $\mathbf{d}^f$ and $\mathbf{h}^f$. The vectors $\mathbf{a}_i$ and $\mathbf{b}_i$ are orthogonal to each other. The maximum number of canonical variate pairs m that can be found is the minimum of the ranks of $\mathbf{d}^f$ and $\mathbf{h}^f$. Since generally $p, q \ll N$ after dimension reduction, $m = min\{p, q\}$. The canonical transformation can be found through the eigen-decomposition of the sample covariance matrix and this CCA transformation is reversible.

If $\mathbf{d}^c$ and $\mathbf{h}^c$ in the canonical space are nearly linearly correlated, a linear model can be used to simulate their relationship. If after CCA, the relationship of $\mathbf{d}^c$ and $\mathbf{h}^c$ is still not quite linear, we can use advanced ML models such as neural networks for regression.

E. Bayesian inference of the prediction

Bayesian inference is widely used for uncertainty quantification because of its conceptual simplicity, mathematical sophistication and rich probabilistic interpretation. In this work, we use the Bayesian inference to estimate predictions. But unlike the traditional workflow which uses Bayesian methods to quantify uncertainties of model parameters first and then infer prediction uncertainties, we use Bayesian methods to calculate the posterior distribution of the predictions directly. Based on Bayes' rule, the posterior distribution of a prediction variable $\mathbf{h}$ for some observed data $\mathbf{d_{obs}}$ is

$$p(\mathbf{h}|\mathbf{d}_{obs}) \propto L(\mathbf{h}|\mathbf{d}_{obs})p(\mathbf{h}), \tag{3}$$

where $p(\mathbf{h})$ is the prior distribution and $L(\mathbf{h}|\mathbf{d}_{obs})$ is the likelihood function. PCA and CCA enable reducing a set of high-dimensional variables $(\mathbf{d}, \mathbf{h})$ to a set of low-dimensional and linearly correlated variables $(\mathbf{d^c}, \mathbf{h^c})$. We first estimate the posterior distribution $p(\mathbf{h}^c|\mathbf{d}_{obs}^c)$ and then transform back $\mathbf{h}^c$ to its original space $\mathbf{h}$. In the canonical space, $p(\mathbf{h}^c|\mathbf{d}_{obs}^c)$ can be estimated by

$$p(\mathbf{h}^c|\mathbf{d}_{obs}^c) \propto L(\mathbf{h}^c|\mathbf{d}_{obs}^c)p(\mathbf{h}^c). \tag{4}$$

We use a linear model G to simulate the relationship between $\mathbf{d}^c$ and $\mathbf{h}^c$, i.e., $\mathbf{d}^c = G\mathbf{h}^c$. By assuming a Gaussian likelihood, $L(\mathbf{h}^c|\mathbf{d}_{obs}^c)$ can be formulated as

$$L(\mathbf{h}^c|\mathbf{d}_{obs}^c) = \exp\Big(-\frac{1}{2}\big(G\mathbf{h}^c - \mathbf{d}_{obs}^c\big)^T \mathbf{C}_{\mathbf{d}^c}^{-1}\big(G\mathbf{h}^c - \mathbf{d}_{obs}^c\big)\Big). \tag{5}$$

Through normal score transformation based on the sample mean $\bar{\mathbf{h}}_{prior}^c$ and the sample covariance $\mathbf{C}_{\mathbf{h}^c}$ calculated from the prior samples of $\mathbf{h}^c$, we obtain a Gaussian prior of $\mathbf{h}^c$ in the transformed space. Since the prior and the likelihood of $\mathbf{h}^c$ are Gaussian, its posterior is also Gaussian and the posterior mean $\tilde{\mathbf{h}}^c$ and posterior covariance $\tilde{\mathbf{C}}_{\mathbf{h}^c}$ can be analytically estimated by

$$\tilde{\mathbf{h}}^c = \bar{\mathbf{h}}_{prior}^c + \mathbf{C}_{\mathbf{h}^c} G^T \big(G\mathbf{C}_{\mathbf{h}^c} G^T + \mathbf{C}_{\mathbf{d}^c}\big)^{-1} \big(\mathbf{d}_{obs}^c - G\bar{\mathbf{h}}_{prior}^c\big), \tag{6}$$

$$\tilde{\mathbf{C}}_{\mathbf{h}^c} = \big(G^T \mathbf{C}_{\mathbf{d}^c}^{-1} G + \mathbf{C}_{\mathbf{h}^c}^{-1}\big)^{-1}, \tag{7}$$

where $\mathbf{C}_{\mathbf{d}^c}$ is the covariance matrix of the observation error. In this work, we are considering a synthetic case where the observed data is one realization from the prior samples, so $\mathbf{C}_{\mathbf{d}^c}$ here is calculated as the covariance of the residuals from the linear model fitting.

An advantage of the Gaussian process regression is that a Gaussian distribution is uniquely defined by its mean and covariance and that sampling a Gaussian distribution is straightforward. Therefore, based on Eq. 6 and 7, we can generate posterior samples of $\mathbf{h}^c$ directly. By undoing the normal score transformation followed by the back transformation of CCA, we obtain posterior samples of $\mathbf{h}^f$. Next, after back transformation of PCA, we obtain the posterior samples of prediction quantity $\mathbf{h}$ in its original space. Then based on these $\mathbf{h}$ samples, we estimate posterior prediction distribution.

III. Application

In this section, we apply LIMP to a regional terrestrial ecosystem model with large uncertain parameters. We evaluate the feasibility and performance of LIMP for effective predictions and rapid data assimilation. For the purpose of proof-of-feasibility, we demonstrate LIMP in a synthetic study where the solution is known and chosen from one of the prior realizations of the prediction variables, and the observed data is the corresponding realization of the observation variables. We evaluate the prediction performance by comparing the results with the reference solution.

A. Description of the terrestrial ecosystem model

The terrestrial ecosystem model (TEM) is developed based on DALEC model [6] and CLM4.5 [7]. The TEM consists of five process-based submodels representing photosynthesis, plant autotrophic respiration, plant carbon allocation, deciduous leaf phenology, and litter and soil organic matter decomposition, to simulate carbon fluxes and state variables using 47 uncertain parameters that are listed in Table I. This TEM includes three vegetation and two soil carbon pools. More information about this model is described in [8]. Here we are interested in simulating annual gross primary productivity (GPP), representing the photosynthetic uptake of carbon by plants, in deciduous forest systems in the eastern region of the United States for 30 years between 1981 and 2010. The region of interest covers 1422 land grid cells (locations) as shown in Fig. 1. As in most land surface models, we assume the 47 model parameters are spatially invariant because the

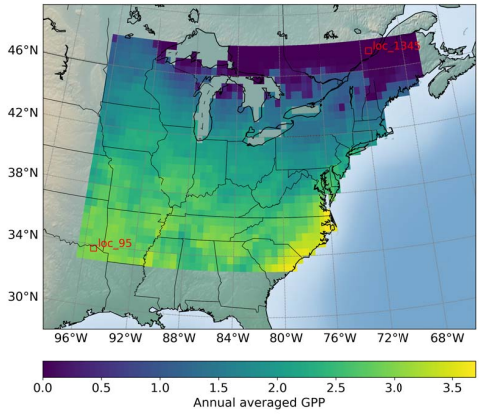

Fig. 1. The simulated region of interest where the GPPs ($gC\ m^{-2}\ d^{-1}$) at the two locations are observation data.

model considers a single plant functional type. Temporal and spatial variations in GPP are driven by changing atmospheric drivers including air temperature, solar radiation, vapor pressure deficit, and CO_2 concentrations that are used as boundary conditions. The carbon state variables are spun up to steady state by cycling the GSWP3 input meteorology [9] from 1981-2010 for five 30-year cycles, and the sixth cycle is used as the output for our LIMP study. One regional TEM run takes about 24 hours on a single processor, which is computationally too expensive for traditional iterative model-data integration studies in 47 parameter space. Here we use LIMP framework to advance model predictions by avoiding the computationally challenging inverse modeling.

We consider two case studies: Case I demonstrates LIMP's capability in efficient model prediction and Case II shows LIMP's capability in fast data assimilation and cost-effective experimental design. In both cases, the prediction QoI is averaged annual GPPs over the entire region between 2001-2010. The observation variables are the annual GPPs between 1981 and 2000 at some locations. In Case I, we use observation data from one location, loc_95 in Fig. 1, to illustrate LIMP and demonstrate its efficient prediction capability. In Case II, we use observation data from two locations, loc_95 and loc_1345, to demonstrate how LIMP can be applied for rapid data assimilation. In both cases, the observation model and the prediction models are the same and they share the same uncertain 47 parameters.

We use high-performance computing to run an ensemble of 2000 TEM simulations in parallel for 30 years between 1981 and 2010 at all locations. The 2000 prior parameter samples are randomly and uniformly drawn from the parameter space defined in Table I, where the parameter ranges are designed to reflect their average values and broad uncertainties associated with the temperate deciduous forest plant functional type. These TEM simulations produce 2000 prior samples of the observation and prediction variables in both Case I and II. We select the first realization of the observation variables as the actual observed data, and the corresponding first realization of the prediction variables as the true prediction value to evaluate the predictive performance. In LIMP, the most computationally expensive part is the forward model simulations which can be run in parallel. After we obtain the prior samples of the observation and prediction variables, we can use data mining techniques and ML methods for efficient model prediction and rapid data assimilation.

TABLE I
MODEL PARAMETERS AND THEIR RANGES TO GENERATE PRIOR SAMPLES.

Parameter	Minimum	Maximum	Units
gdd_crit	100	700	$degC\ day$
$crit\_dayl$	36000	43000	$seconds$
$ndays\_on$	15	45	$days$
$ndays\_off$	7.5	22.5	$days$
nue_tree	7.5	37.5	$none$
nue_grass	4.5	13.5	$none$
$slatop\_everg$	0.005	0.015	m^2/gC
$slatop\_decid$	0.015	0.045	m^2/gC
$livewdcn$	25	75	gC/gN
$leafcn$	12.5	37.5	gC/gN
$frootcn$	21	63	gC/gN
$fstor2tran$	0.25	0.75	$none$
$stem\_leaf$	1.35	4.05	$none$
$croot\_stem$	0.15	0.45	$none$
f_livewd	0.05	0.15	$none$
$froot\_leaf$	0.5	1.5	$none$
rg_frac	0.15	0.45	$none$
br_mr	1.26×10^{-6}	3.78×10^{-6}	$gC/m^2/s/gN$
$q10\_mr$	0.75	2.25	$none$
$cstor\_tau$	1.5	4.5	$year$
r_mort	0.1	0.3	$1/year$
$lwtop\_ann$	0.35	1.05	$1/year$
$leaf\_long$	1.5	4.5	$year$
$froot\_long$	1.5	4.5	$year$
$q10\_hr$	0.75	2.25	$none$
k_l1	0.602	1.806	$1/day$
k_l2	0.036	0.109	$1/day$
k_l3	0.007	0.02	$1/day$
k_s1	0.036	0.109	$1/day$
k_s2	0.007	0.021	$1/day$
k_s3	7.0×10^{-4}	2.1×10^{-3}	$1/day$
k_s4	5.0×10^{-5}	1.5×10^{-4}	$1/day$
k_frag	5.0×10^{-4}	1.5×10^{-3}	$1/day$
rf_l1s1	0.19	0.57	$none$
rf_l2s2	0.275	0.825	$none$
rf_l3s3	0.145	0.435	$none$
rf_s1s2	0.14	0.42	$none$
rf_s2s3	0.23	0.69	$none$
rf_s3s4	0.275	0.825	$none$
$soil4ci$	500	1500	gC/m^2
cwd_flig	0.12	0.36	$none$
fr_flig	0.125	0.375	$none$
lf_flig	0.125	0.375	$none$
fr_flab	0.125	0.375	$none$
lf_flab	0.125	0.375	$none$
fpi	0.05	0.15	$none$
fpg	0.7	0.95	$none$

B. Case study I: efficient and effective model prediction

Fig. 2 shows the 2000 prior samples of the annual GPPs at loc_95 (variables $\mathbf{d}$) and the averaged annual GPPs over the entire region (variables $\mathbf{h}$). Based on these prior samples, we simulate the relationship between $\mathbf{d}$ and $\mathbf{h}$ from which we use the observed data ($\mathbf{d}_{obs}$, the red line in Fig. 2(a)) to predict $\mathbf{h}$ whose reference value is shown in the red line in

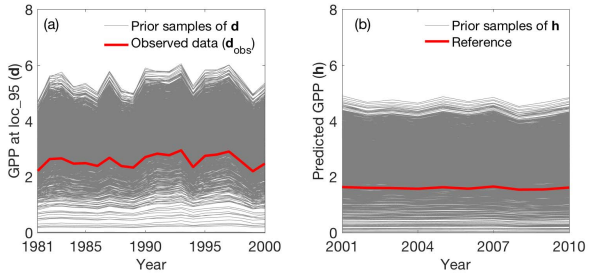

Fig. 2. Prior samples of observation variables $\mathbf{d}$ and prediction variables $\mathbf{h}$.

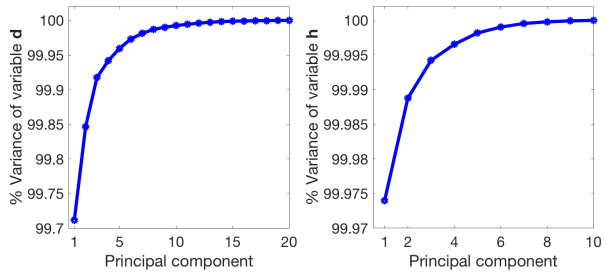

Fig. 3. Scree plots of observation variables $\mathbf{d}$ and prediction variables $\mathbf{h}$.

Fig. 2(b). The observation variable $\mathbf{d}$ is a time series having 20 elements (dimensions) representing annual GPPs from 1981-2000 at the sample location and the prediction variable $\mathbf{h}$ is another time series having 10 dimensions representing the annual GPPs from 2001-2010 averaged over the domain. We use FPCA to reduce the dimensions of $\mathbf{d}$ and $\mathbf{h}$ to remove the multicollinearity and facilitate the establishment of their statistical relationship. The scree plot shown in Fig. 3 indicates that only the first one principal component is required to capture over 99% of the variation in both the observation and prediction variables. This suggests that FPCA effectively reduces the 20 dimensions of $\mathbf{d}$ into one dimension and the 10 dimensions of $\mathbf{h}$ into one dimension. The large dimension reduction is attributed to the fact that the observation and prediction variables are annual GPPs which exhibit smooth changes along years (Fig. 2). Now we establish the statistical relationship of $\mathbf{d}$ and $\mathbf{h}$ in their reduced dimensions, namely, $\mathbf{d}^f$ and $\mathbf{h}^f$.

The scatter plot of Fig. 4(a) indicates that $\mathbf{d}^f$ and $\mathbf{h}^f$ have strong linear correlation with coefficient over 0.99. This suggests that a linear regression model can be established to simulate the relationship of $\mathbf{d}^f$ and $\mathbf{h}^f$. In this case study, both observation and prediction variables are the same type of quantity with smooth variation, so it is not surprising that they show strong linear correlation. Practically this situation is not very rare, for example, we use historical data to predict the future behavior at the same location, and use data from one location to predict the same variable at other correlated locations. However, in most situations, a nonlinear relationship between $\mathbf{d}^f$ and $\mathbf{h}^f$ is expected and then CCA is necessary for linearization. Fig. 4(b) indicates that the prior samples

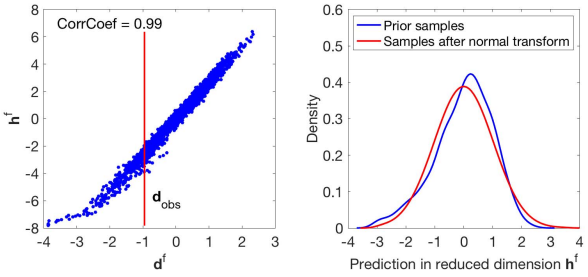

Fig. 4. (a) Scatter plots of $\mathbf{d}^f$ and $\mathbf{h}^f$ in reduced dimension; (b) Estimated probability density function of prior samples and the samples after normal score transformation of $\mathbf{h}^f$.

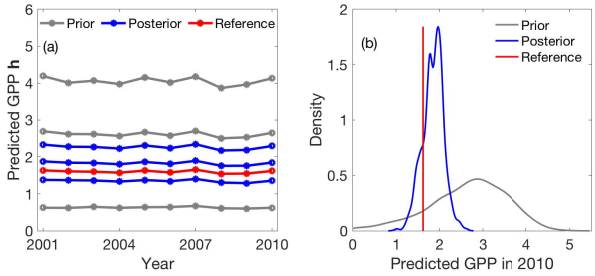

Fig. 5. Prior and posterior predictions with the reference in case study I; in (a) the middle gray/blue line represents prior/posterior mean and the top and bottom gray/blue lines quantify 95% prior/posterior confidence interval.

of $\mathbf{h}^f$ are not Gaussian distributed. We thus apply a normal score transformation based on their sample mean and sample variance and make the prior distribution of $\mathbf{h}^f$ Gaussian as shown in Fig. 4(b).

By assuming a Gaussian likelihood, with the Gaussian prior and a linear relationship between $\mathbf{d}^f$ and $\mathbf{h}^f$, the posterior distribution $p(\mathbf{h}^f|\mathbf{d}^f_{obs})$ is also Gaussian with mean and covariance calculated in Eq. 6 and 7. Next, we draw posterior samples from this Gaussian distribution. Last, we do a series of back transformation to transform those posterior samples in the reduced space $\mathbf{h}^f$ back into their originally high-dimensional space $\mathbf{h}$. We start by undoing the normal score transformation, followed by the FPCA back transformation. Note that in this case study, we do not need CCA for linearization. If CCA is used, we need to undo the canonical transformation before the FPCA back transformation. The final prediction results of $\mathbf{h}$ are summarized in Fig. 5. Fig. 5(a) shows the means and 95% confidence intervals of the prior and posterior samples. Fig. 5(b) plots the prior and posterior probability density function of the predicted GPPs in year 2010. The figures indicate that the posterior estimation improves prediction accuracy and reduces predictive uncertainty in comparison to the prior. The root mean squared error (RMSE) of the posterior mean is 0.23, five times smaller than that of the prior mean of 1.02. The posterior uncertainty bound encloses the reference with a much narrower averaged range of 0.91 compared to that of 3.42 of the prior.

This study of Case I shows that based on 2000 parallel

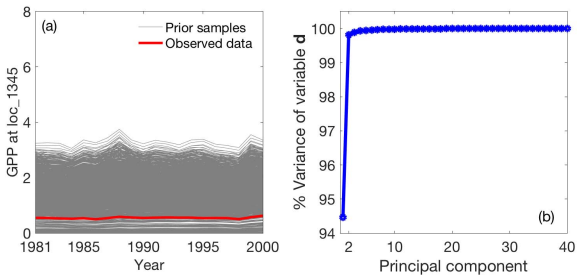

Fig. 6. (a) Prior samples at location 1345; (b) scree plot of observation variables at locations 95 and 1345 after FPCA and mixed PCA.

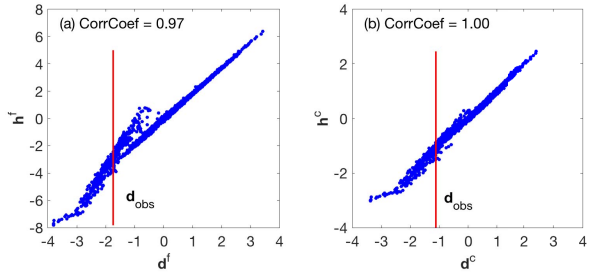

Fig. 7. (a) Scatter plots of $\mathbf{d}^f$ and $\mathbf{h}^f$ in reduced dimension after PCA; (b) scatter plots of $\mathbf{d}^c$ and $\mathbf{h}^c$ in canonical space after CCA.

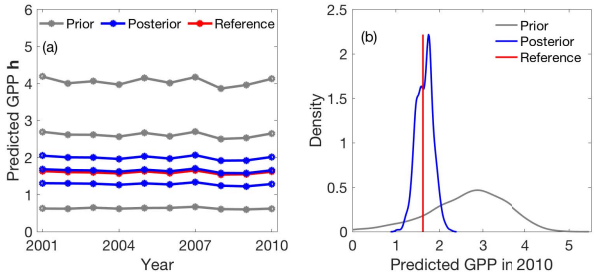

Fig. 8. Prior and posterior predictions with the reference in case study II; in (a) the middle gray/blue line represent prior/posterior mean and the top and bottom gray/blue lines quantify 95% prior/posterior confidence interval.

model evaluations, our LIMP framework can produce an accurate and credible prediction of a complex ecosystem model in a Bayesian context, which demonstrates the LIMP's capability of efficient and effective model prediction. In fact, as the observation and prediction responses in this work are rather smooth and their prior samples include the actual observed data, using 1000 samples we can get the similar prediction results as the 2000 samples. The traditional iterative inversion-based method is not a good solution for this problem. First, this model has more parameters than the observation data which requires solving an ill-posed inverse problem. Second, with this small number of model evaluations, it is hardly possible for the traditional iterative method to get a reasonable parameter calibration in the 47 parameter space, not even mention that the traditional method cannot be fully parallel and additionally needs the forward model simulations for the predictions. Furthermore, the surrogate modeling cannot address these challenges because of the high parameter dimensionality and that we still need solve the under-determined inverse problem in a likely inaccurate surrogate space.

C. Case study II: fast data assimilation and experimental design

In this case study II, we consider observation data from one more location besides the loc_95 considered in Case I and demonstrate how LIMP can be used for fast assimilation of additional data and for experimental design to guide new data collection. The prediction quantity is the same as the Case I, so the information of $\mathbf{h}^f$ in the reduced dimension can be reused. The additional observation data is 20 years annual GPPs from loc_1345 (Fig. 1). The 2000 prior samples of this observation variable are shown in Fig. 6(a) together with its observed values. Now our observation variables contain two 20-dimensional time series. We use the mixed PCA to compress the 40-dimensional variables and Fig. 6(b) indicates that the first two principal components are required to capture over 99% of the variation in the simulated observations. Fig. 7(a) uses the scatter plot to analyze the relationship between $\mathbf{d}^f$ and $\mathbf{h}^f$ in the reduced dimensions and indicates that although $\mathbf{d}^f$ and $\mathbf{h}^f$ has a high correlation coefficient of 0.97, their correlation is not quite linear. Two disconnected patterns can be clearly identified from the scatter points and the linear regression is not expected to perform well. We then apply the CCA to transform $\mathbf{d}^f$ and $\mathbf{h}^f$ into a canonical space, and their corresponding samples of $\mathbf{d}^c$ and $\mathbf{h}^c$ in Fig. 7(b) indicates that $\mathbf{d}^c$ and $\mathbf{h}^c$ has a strong linear correlation with a coefficient nearly 1.0. This suggests that a linear regression model can be established to simulate the observation-prediction relationship in the transformed space. Then the following procedure is similar as the Case I, i.e., using Eq. 6 and 7 to draw Gaussian posterior samples of $\mathbf{h}^c$ and then transforming them back into posterior samples of $\mathbf{h}$ in the original space through a series of back transformation, starting from the normal score back transformation, then the canonical back transformation, and last the mixed PCA back transformation.

The final prediction results of $\mathbf{h}$ are summarized in Fig. 8. The figure indicates that comparing to Fig. 5 in Case I, adding one more observation data further improves the prediction accuracy and reduces the predictive uncertainty. As shown in Fig. 8(a), the posterior mean based on two observation sets is closer to the reference with the RMSE of 0.047 and the posterior uncertainty bound snugly encloses the reference with the averaged range of 0.71. Fig. 8(b) shows that the posterior density function can predict the reference with a very high probability.

This study of Case II indicates that our LIMP framework can effectively assimilate additional data to improve predictions. This additional data does not necessarily have to come from other locations, it can also be measurements over a longer period of time from the same location, or other types of measurements. Moreover, assimilation of additional data in

LIMP is very computationally efficient. It does not cost much computing compared to the single dataset case, as long as we have the prior samples of the additional data simulated. The capability of LIMP in efficient and effective data assimilation is promising for the experimental design. When the prediction based on existing observation data is poor, we usually want to collect more data to improve the predictive performance. However, designing a cost-effective data collection scheme is challenging because of the large search space, multiple uncertainty sources, and massive computational costs [10]. LIMP allows optimization of an experimental design in an efficient way with considering multiple sources of uncertainties in a Bayesian framework.

IV. Discussions and conclusions

Model-data integration problems in Earth system modeling are generally formulated as iterative model inversion problems that are computationally expensive and challenging. Although model calibration is important for understanding processes and effects within the Earth system, when the concern is about improving model prediction rather than obtaining a well-calibrated model, the proposed LIMP framework has outstanding advantages in computational efficiency. LIMP establishes a statistical relationship between the observation and prediction variables and infers Bayesian prediction probability distribution based on actual observed data. Thus, LIMP is particularly suitable for Earth science problems where the models have a large number of uncertain parameters with heterogeneous uncertainties, which makes traditional model inversion infeasible, and where the observation and prediction data are usually time series or spatial response maps, which promises an easy establishment of observation-prediction relationship after dimension reduction.

In this effort, we demonstrate the capabilities of LIMP in efficient and effective model prediction and in fast data assimilation using a regional terrestrial ecosystem model with 47 parameters. The results indicate that LIMP improves the prediction accuracy and reduces the predictive uncertainty in estimating the regional GPPs. And the computational cost to implement the LIMP is 2000 forward model simulations which can be parallel, although we can use 1000 simulations to get the similar results. In addition, LIMP allows quick assimilation of the additional data to continually improve the predictive performance without extra computing efforts, as long as the prior samples of the additional data are already generated. Compared to the traditional inversion-based model prediction, the computational saving of LIMP can be outstanding. To calibrate a complex ecosystem model with 47 parameters, typically $\sim 10^5$ iterative model runs are needed, while with supercomputers, LIMP can be evaluated using the computing time of one forward model run.

On the other hand, since LIMP uses estimated observation-prediction relationship to infer predictions based on the actual observed data, establishing the statistical regression model is crucial. In this work, we use PCA followed by CCA to build a linear relationship in the reduced canonical space and then use the Gaussian linear regression for predictions. In situations when the relationship is nonlinear and multimodal even after PCA and CCA transformations, we can use advanced machine learning models such as Bayesian neural networks for regression. In addition, also important for accurate predictions from LIMP is that the prior samples should enclose the actual observed data, otherwise extrapolation may occur. We can first use outlier detection techniques such as support vector machine to identify whether the prior is consistent with the observation data. If not, we may increase the prior sample size, enlarge the prior uncertainty bound, use a different prior, or consider multiple models to ensure that the observation data lies inside of the span of the prior samples. Another scenario that could affect the LIMP performance is that there are insufficient samples in vicinity of the observation data even though it is covered by the prior samples. This could happen when the observation data lies in the extremes of the prior and there are not enough samples to infer a good posterior distribution of the prediction. This problem can be solved using importance sampling to purposely generate more samples around the observation data. Last, note that although LIMP uses statistical methods to infer predictions, it values the physics and processes to simulate the Earth system. LIMP learns observation-prediction relationship based on model simulation samples. If the Earth system model cannot predict future system behavior well, then LIMP may not give a reasonable prediction. In this sense, LIMP can help fast diagnose model structural errors and guide model development and improvment.

Although some factors may impair its performance, the LIMP framework will have considerable impact on Earth system modeling because of its computational advantages over the traditional inversion-based modeling. An additional appealing characteristic of applying LIMP in Earth sciences is that as long as the relevant model outputs are saved, LIMP can be implemented using existing ensembles of Earth system model simulations rather than requiring new computationally expensive simulations.

Acknowledgment

Primary support for this work was provided by the Scientific Discovery through Advanced Computing (SciDAC) program, funded by the U.S. Department of Energy (DOE), Office of Advanced Scientific Computing Research (ASCR) and Office of Biological and Environmental Research (BER). Additional support was provided by BER's Terrestrial Ecosystem Science Scientific Focus Area (TES-SFA) project and Oak Ridge National Laboratory (ORNL) AI initiative project. The authors are supported by ORNL, which is supported by the DOE under contract DE-AC05-00OR22725.

References

[1] A. Fox, et al., "The REFLEX project: Comparing different algorithms and implementations for the inversion of a terrestrial ecosystem model against eddy covariance data", Agric. For. Meteorol., 149, 1597-1615, 2009.

[2] D. Lu, D. Ricciuto, A. Walker, C. Safta, and W. Munger, "Bayesian calibration of terrestrial ecosystem models: a study of advanced Markov chain Monte Carlo methods", Biogeosciences, 14, 4295-4314, 2017.
[3] J. Ray, Z. Hou, M. Huang, K. Sargsyan, and L. Swiler, "Bayesian calibration of the Community Land Model using surrogates", SIAM/ASA J. Uncertain. Quantif., 3, 199-233, 2015.
[4] S. Razavi, B. A. Tolson, and D. H. Burn, "Review of surrogate modeling in water resources", Water Resour. Res., 48, W07401, 2012.
[5] W. J. Krzanowski, "Principles of Multivariate Analysis: A User's Perspective", Oxford Stat. Ser. 22, revised ed., Oxford Univ. Press, N. Y., 2000.
[6] M. Williams, P. A. Schwarz, B. E. Law, J. Irvine, and M. Kurpius, "An improved analysis of forest carbon dynamics using data assimilation", Glob. Change Biol., 11, 89-105, 2005.
[7] K. W. Oleson, and D. M. Lawrence, "Technical description of version 4.5 of the Community Land Model (CLM)", NCAR Tech. Note NCAR/TN-5031STR, 420 pp., National Center for Atmospheric Research, Boulder, CA, USA, https://doi.org/10.5065/D6RR1W7M, 2013.
[8] D. Lu, and D. Ricciuto, "Efficient surrogate modeling methods for large-scale Earth system models based on machine-learning techniques", Geosci. Model Dev., 12, 1791-1807, 2019.
[9] H. Kim, "Global soil wetness project phase 3 atmospheric boundary conditions (experiment 1), Data Integration and Analysis System (DIAS)", https://doi.org/10.20783/DIAS.501, 2017.
[10] D. Lu, M. Ye, S. P. Neuman, and L. Xue, "Multimodel Bayesian analysis of data-worth applied to unsaturated fractured tuffs", Adv. in Water Resour., 35, 69-82, 2012.