SIAM/ASA J. Uncertainty Quantification
Vol. 8, No. 3, pp. 1118–1138

Resource-Constrained Model Selection for Uncertainty Propagation and Data Assimilation*

Lun Yang†, Peng Wang‡, and Daniel M. Tartakovsky§

Abstract. All observable phenomena can be described by alternative mathematical models, which vary in their fidelity and computational cost. Selection of an appropriate model involves a tradeoff between computational cost and representational accuracy. Ubiquitous uncertainty (randomness) in model parameters and forcings, and assimilation of observations of the system states into predictions, complicate the model selection problem. We present a framework for analysis of the impact of data assimilation on cost-constrained model selection. The framework relies on the definitions of cost and accuracy functions in the context of data assimilation for multifidelity models with uncertain (random) coefficients. It contains an estimate of error bounds for a system's state prediction obtained by assimilating data into a model via an ensemble Kalman filter. This estimate is given in terms of model error, sampling error, and data error. Two examples illustrating the applicability of our model selection method are provided. The first example deals with an ordinary differential equation, for which a sequence of lower-fidelity models is constructed by progressively increasing the time step used in its discretization. The second example comprises the viscous Burgers equation as the high-fidelity model and a linear advection-diffusion equation as its low-fidelity counterpart.

Key words. multifidelity models, data assimilation, model selection, ensemble Kalman filter, uncertainty quantification, Monte Carlo

AMS subject classifications. 60H30, 35Q86, 76S05, 86A05

DOI. 10.1137/19M1263376

1. Introduction.

All observable phenomena can be described by alternative mathematical models, which vary in their fidelity and computational cost. Invariably, high-fidelity models have higher computational cost than their low-fidelity counterparts. Selection of an appropriate model involves a tradeoff between computational cost and representational accuracy. In most applications it is common to choose as "accurate" a model (more physics, more degrees of freedom) as the available computational resource allows.

Yet, ubiquitous uncertainty (randomness) in model parameters and forcings complicates the selection problem. That is because solutions of models with random coefficients are given

*Received by the editors May 22, 2019; accepted for publication (in revised form) June 30, 2020; published electronically August 20, 2020.

https://doi.org/10.1137/19M1263376

Funding: The work of the first two authors was partially funded by National Key Research and Development Program of China (grant 2017YFB0701700) and (grant 2018YFB0703902); L. Yang's stay at Stanford was funded by the China Scholarship Council Foundation. The work of the third author was supported in part by Air Force Office of Scientific Research under award FA9550-17-1-0417, by U.S. Department of Energy under award DE-SC0019130, and by a gift from TOTAL.

†LMIB & School of Mathematical Sciences, Beihang University, Beijing, 100191, China (lun.yang@buaa.edu.cn).

‡LMIB & School of Mathematical Sciences, School of Microelectronics, Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University, Beijing 100191, China (wang.peng@buaa.edu.cn).

§Department of Energy Resources Engineering, Stanford University, Stanford, CA 94305 (tartakovsky@stanford.edu).

in terms of their probability density functions (PDFs) or, in many practical settings, in terms of their statistical moments such as ensemble means and variances. More often than not, these are computed with sampling methods, e.g., Monte Carlo simulations [20] or stochastic collocation techniques [11]. When that happens, the solution error $\mathcal{E} = \mathcal{E}^{\text{rep}} + \mathcal{E}^{\text{sam}}$ consists of the representational error $\mathcal{E}^{\text{rep}}$ that reflects a model's fidelity and the sampling error $\mathcal{E}^{\text{sam}}$ due to a finite/small number of samples used to estimate a solution's statistics. Given limited computational resources, $\mathcal{E}^{\text{sam}}$ goes up as $\mathcal{E}^{\text{rep}}$ goes down since one can compute more realizations of a low-fidelity model than of a high-fidelity model during an allocated simulation time.

Availability of observations of the system states adds a further complication to the model selection process. PDFs of solutions of multifidelity models can now be thought of as prior distributions that are refined with data. Given a sufficient amount of observations, the resulting posterior distribution of a low-fidelity model solution can become as accurate as that of a high-fidelity model solution. That would render the reliance on the high-fidelity model superfluous. Of direct relevance to our study are ensemble-based data assimilation techniques, such as Bayesian update, particle filters, and ensemble Kalman filters (EnKFs). Similar to the methods of forward uncertainty propagation discussed above, their performance is expected to degrade as the number of model runs one can afford decreases due to the model's complexity. This interplay of data and limited computational resources was studied via numerical experimentation in the context of multifidelity models of multiphase flow in porous media [17]. The goal of our investigation is to provide a theoretical foundation for resource-constrained model selection in the presence of data.

This aim constitutes another facet of multifidelity simulations in the context of computational resource-constrained model selection. It differs from Bayesian model selection [3, 4] and Bayesian model averaging [6, 10], both of which deal with multiple models whose relative veracity is a priori unknown. It is also distinct from data assimilation with multiple models, which combines models of various fidelity either to maximize the predictive accuracy [14, 21] or to minimize the computational cost while maintaining a prescribed accuracy [9, 13, 16] through, e.g., multilevel [8] or multifidelity [15] Monte Carlo sampling. Finally, it does not use model complexity as an argument for model selection [12]. Instead, our analysis contributes to the ongoing debate of whether, and under what conditions, practical constraints of available computing time and uncertain input parameters warrant the use of more sophisticated numerical models.

In section 2 we formulate a data assimilation problem in the context of multifidelity simulations. Section 3 contains definitions of the cost and accuracy functions and their use for the resource-constrained selection between high- and low-fidelity models in the absence of system state observations. The impact of data on such a selection is investigated in section 4 by analyzing the accuracy and convergence properties of EnKF. Our theoretical criteria for the resource-constrained model selection in the presence of noisy data are verified in section 5 by using an ordinary differential equation and the viscous Burgers equation as examples. Major conclusions drawn from our study are summarized in section 6.

2. Problem formulation. Consider a system state $v \in \mathbb{R}^{N_v}$ $(N_v \geq 1)$ that is represented by the time sequence $\{v_i\}_{i\in\mathbb{N}}$ forming a Markov chain. Given the initial state v_0, the operator $\Psi : \mathbb{R}^{N_v} \to \mathbb{R}^{N_v}$ describes the true temporal evolution of v, from time t_i to time t_{i+1},

$$v_{i+1} = \Psi(v_i), \qquad i \in \mathbb{N}. \tag{1}$$

Uncertainty in the initial state v_0 is treated by representing the latter as a Gaussian random field with the mean $\mu_0 \in \mathbb{R}^{N_v}$ and the real covariance matrix $C_0 \in \mathbb{R}^{N_v \times N_v}$. The unknown (and unknowable) operator Ψ is replaced with its high-fidelity counterpart Ψ^{h} and with a set Ψ^{l} of m low-fidelity counterparts, $\mathrm{l} = \{l_1, \ldots, l_m\}$, all of which are continuous in $\mathbb{R}^{N_v} \times \mathbb{R}^{N_v}$. The corresponding model errors ξ^{h} and ξ^{l} are described by independently and identically distributed (i.i.d.) (white noise) sequences $\{\xi_i^{\mathrm{h}}\}_{i\in\mathbb{N}}$ and $\{\xi_i^{\mathrm{l}}\}_{i\in\mathbb{N}}$ with a given distribution, e.g., $\xi_i^k \sim \mathcal{N}(0, \Sigma^k)$ with $k = \mathrm{h}$ or l. (More sophisticated representations of the model error can be found in, e.g., [7, section 4.2.1] and can be readily incorporated into the present analysis.) Then, predictions of the high-fidelity ($v^{\mathrm{h}} \in \mathbb{R}^{N_v}$) and low-fidelity ($v^{\mathrm{l}} \in \mathbb{R}^{N_v}$) models satisfy

$$\hat{v}_{i+1} = \Psi^k(\hat{v}_i) + \xi_i^k, \qquad k = \mathrm{h} \text{ or } \mathrm{l}, \tag{2}$$

respectively. These equations are subject to the random initial conditions $\hat{v}_0 \sim \mathcal{N}(\mu_0, C_0)$.

The high- and low-fidelity models (2) are supplemented with noisy (possibly indirect) N_y-dimensional observations $y_i \in \mathbb{R}^{N_y}$. At any time t_i, these are related to the state variable v_i by the observation operator $h : \mathbb{R}^{N_v} \to \mathbb{R}^{N_y}$ such that

$$y_i = h(v_i) + \eta_i, \tag{3}$$

where the measurement noise $\{\eta_i\}_{i\in\mathbb{N}}$ is represented by an i.i.d. sequence of Gaussian random variables, $\eta_i \sim \mathcal{N}(0, \Gamma)$. Note that the Gaussianity assumption for the model and data noise is made for the sake of concreteness only. We use EnKF to assimilate the data (3) into the low- and high-fidelity model (2).

Our study deals with two issues: the role played by limited computational resources on the selection between the high- and low-fidelity models (section 3) and the impact of observations on this tradeoff (section 4).

3. Cost-accuracy tradeoff in model selection.

We start by considering, in subsection 3.1, Monte Carlo simulations for forward uncertainty propagation in multifidelity models. We present definitions of the cost and accuracy functions in the context of multifidelity data assimilation in subsection 3.2. These definitions are used in subsection 3.3 to demonstrate the relative accuracy of high- and low-fidelity models when the computational cost is fixed. Our strategy for selecting a model for uncertainty propagation under resource constraints is presented in subsection 3.4.

3.1. Model prediction via Monte Carlo sampling.

In the absence of observational data, a model is the only means of estimation of a system state. The true PDF $f_{\hat{v}_i}$ of the random state $\hat{v}_i$ in (2) is approximated by the sample estimate $\hat{f}_{\hat{v}_i}$, computed from N samples of solutions of (2) and each corresponding to a sample of the initial condition $\hat{v}_0^{(n)}$ drawn from $\mathcal{N}(\mu_0, C_0)$,

$$\hat{v}_{i+1}^{(n)} = \Psi^k\left(\hat{v}_i^{(n)}\right) + \xi_i^{k(n)}, \qquad n = 1, \ldots, N. \tag{4}$$

To simplify the notation, we consider a single state variable ($N_v = 1$), whose sample mean and variance,

$$\hat{\mu}_i = \frac{1}{N}\sum_{n=1}^{N} \hat{v}_i^{(n)} \quad \text{and} \quad \hat{\sigma}_i^2 = \frac{1}{N-1}\sum_{n=1}^{N}\left(\hat{v}_i^{(n)} - \hat{\mu}_i\right)^2, \tag{5}$$

approximate the ensemble mean and variance,

$$\mu_i \equiv \mathbb{E}(\hat{v}_i) = \int_{\mathbb{R}} v_i' \hat{f}_{\hat{v}_i}(v_i')\mathrm{d}v_i' \quad \text{and} \quad \sigma_i^2 \equiv \mathbb{E}\left[(\hat{v}_i - \mu_i)^2\right] = \int_{\mathbb{R}} (v_i' - \mu_i)^2 \hat{f}_{\hat{v}_i}(v_i')\mathrm{d}v_i'.$$

Let the true system state v satisfy

$$v_{i+1} = \Psi(v_i), \qquad v_0 = u. \tag{6}$$

Then, the absolute error $\mathcal{E}_{i+1} = |\hat{\mu}_{i+1} - v_{i+1}|$ is bounded by the triangle inequality

$$\begin{aligned} \mathcal{E}_{i+1} &= |\hat{\mu}_{i+1} - \mathbb{E}[\Psi^k(\hat{v}_i)] + \mathbb{E}[\Psi^k(\hat{v}_i)] - \Psi^k(v_i) + \Psi^k(v_i) - \Psi(v_i)| \\ &\le \underbrace{|\hat{\mu}_{i+1} - \mathbb{E}[\Psi^k(\hat{v}_i)]|}_{\mathcal{E}_i^{\mathrm{sam}}} + \underbrace{|\mathbb{E}[\Psi^k(\hat{v}_i)] - \Psi^k(v_i)|}_{\mathcal{E}_i^{\mathrm{ini}}} + \underbrace{|\Psi^k(v_i) - \Psi(v_i)|}_{\mathcal{E}_i^{\mathrm{rep}}}. \end{aligned} \tag{7}$$

According to the "rule of three sigmas," $\mathbb{P}[\mathcal{E}_i^{\mathrm{sam}} < 3\sigma_{i+1}/\sqrt{N}] \approx 0.997$, even though practical computations often characterize the sampling error $\mathcal{E}_i^{\mathrm{sam}}$ in terms of the probable error $0.6745\sigma_{i+1}/\sqrt{N}$, e.g., [18]. More generally, the sampling error $\mathcal{E}_i^{\mathrm{sam}}$ decreases with the inverse of the power of the number of samples, $\mathcal{E}_i^{\mathrm{sam}} \sim 1/N^{\alpha_1}$; in the case of Monte Carlo simulations described in the previous section, $\alpha_1 = 1/2$, the value we employ below. The model error $\mathcal{E}_i^{\mathrm{rep}}$ increases with the distance between the the high- or low-fidelity model Ψ^k and its true counterpart Ψ, the fact that we codify with a relation $\mathcal{E}_i^{\mathrm{rep}} \sim |\Psi - \Psi^k|^{\beta_1}$ in which we set $\beta_1 = 1$. Implicit in the latter assumption is the notion that the prediction uncertainty (variance of the predicted system state, C_i) of the high- and low-fidelity models is approximately the same. This assumption is not universally valid, e.g., a high-fidelity model ($k = \mathrm{h}$) might have more uncertain parameters that give rise to larger prediction variance C_i.

We define the sampling error $\mathcal{E}^{\mathrm{sam}}$, the representational or model error $\mathcal{E}^{\mathrm{rep}}$, and the initialization error $\mathcal{E}^{\mathrm{ini}}$ of a model prediction as the suprema of their corresponding errors at each time step,

$$\mathcal{E}^{\mathrm{sam}} \equiv \sup_{i\in\mathbb{N}} \mathcal{E}_i^{\mathrm{sam}} = \frac{c_1^k}{\sqrt{N}}, \quad \mathcal{E}^{\mathrm{rep}} \equiv \sup_{i\in\mathbb{N}} \mathcal{E}_i^{\mathrm{rep}} = c_2^k|\Psi - \Psi^k|, \quad \mathcal{E}^{\mathrm{ini}} \equiv \sup_{i\in\mathbb{N}} \mathcal{E}_i^{\mathrm{ini}} = c_3^k|v_0 - \mu_0|, \tag{8}$$

where $|\Psi - \Psi^k| = \max_{i\in\mathbb{N}} |\Psi(v_i) - \Psi^k(v_i)|$ and c_3^k depends on the Lipschitz continuity assumption of Ψ^k. The total prediction error $\mathcal{E}$, defined as the upper bound on the error at any given time over the sequence, is the sum of these three errors,

$$\mathcal{E} = \mathcal{E}^{\mathrm{sam}} + \mathcal{E}^{\mathrm{rep}} + \mathcal{E}^{\mathrm{ini}}. \tag{9}$$

To focus on the effects of a finite number of samples N and model fidelity Ψ^k, we neglect the possible discrepancy between the mean of the initial state μ_0 and its true value u_0, i.e., set the initialization error to $\mathcal{E}^{\mathrm{ini}} = 0$. With these assumptions, an estimate of the solution error becomes

$$\mathcal{E}(\Psi^k, N) = c_1^k/\sqrt{N} + c_2^k|\Psi - \Psi^k|, \qquad k = \mathrm{h} \text{ or } \mathrm{l}. \tag{10}$$

3.2. Cost and accuracy functions. Let $\mathcal{C} = \mathcal{C}(\Psi^k, N)$ denote the cost associated with computing N realizations of the model of fidelity k ($k = \text{h or l}$). For a given N,

$$\mathcal{C}(\Psi^{\text{h}}, N) \geq \mathcal{C}(\Psi^{\text{l}}, N) > 0, \tag{11}$$

with the equality achieved when the high- and low-fidelity models coincide, $\Psi^{\text{h}} = \Psi^{\text{l}}$. Likewise, the cost of the low-fidelity models introduced in section 2 satisfies the inequalities $\mathcal{C}(\Psi^{l_1}, N) \geq \mathcal{C}(\Psi^{l_2}, N) \geq \cdots \geq \mathcal{C}(\Psi^{l_m}, N)$. The cost $\mathcal{C}$ increases with N and decreases with the distance between the true model Ψ and its approximation Ψ^k. The latter postulate codifies a typical feature of multifidelity models: the higher a model's fidelity, the closer it is to the "true" model (by definition) and the more computationally expensive it becomes. For $k = \text{h or l}$, we take this relationship to be of the form $\mathcal{C}(\Psi^k, N) \sim N^\alpha/|\Psi - \Psi^k|^\beta + \gamma^k$, where γ^k represents the overhead cost of the analysis of the kth model. Setting the exponents to $\alpha = 1$ and $\beta = 1$ and the overhead cost to $\gamma^k = 0$ for the sake of concreteness, this yields

$$\mathcal{C}(\Psi^k, N) = c_0^k \frac{N}{|\Psi - \Psi^k|}, \qquad k = \text{h or l}, \tag{12}$$

where $c_0^k \in \mathbb{R}^+$ is the constant of proportionality.

To simplify the subsequent analysis, we assume the proportionality constants $c_0, c_1, c_2 \in \mathbb{R}^+$ to be independent of the model fidelity. The assumption of the fidelity-independent constant c_2 suggests the existence of a hierarchy of high- and low-fidelity models, as demonstrated by the hierarchical time-stepping model of an ODE in subsection 5.1. This assumption is not universally valid, as illustrated by a viscous Burgers equation and its low-fidelity linearized counterpart in subsection 5.2. The general case of the model-dependent constants c_0^k, c_1^k, and c_2^k is treated in Appendix A.

Given a fixed cost $\mathcal{C}_0$, it follows from (12) that $|\Psi - \Psi^k| = c_0 N^k/\mathcal{C}_0$, where N_k is the maximum affordable number of samples of the kth model. Then, (10) yields a dependence of the simulation error $\mathcal{E}$ on N^k for a given cost $\mathcal{C}_0$,

$$\mathcal{E}(\cdot, N^k) = \frac{c_1}{\sqrt{N^k}} + \frac{c_0 c_2}{\mathcal{C}_0} N^k, \qquad k = \text{h or l}. \tag{13}$$

The identical procedure used to eliminate N^k from (10) gives

$$\mathcal{E}(\Psi^k, \cdot) = \sqrt{\frac{c_0}{\mathcal{C}_0}} \frac{c_1}{\sqrt{|\Psi - \Psi^k|}} + c_2 |\Psi - \Psi^k|, \qquad k = \text{h or l}. \tag{14}$$

The dependence (13) is shown in Figure 1 for $c_0 = 1, c_1 = 15, c_2 = 1$, and $\mathcal{C}_0 = 10$. When one can afford but a small number of Monte Carlo runs, the total error $\mathcal{E}$ is dominated by the sampling error $\mathcal{E}^{\text{sam}}$ which decreases with N^k. As N^k becomes sufficiently large, the model error $\mathcal{E}^{\text{rep}}$ dominates the total error $\mathcal{E}$; given the fixed computational cost $\mathcal{C}_0$, and within the framework established by (12), the model discrepancy $|\Psi - \Psi^k| \sim N^k/\mathcal{C}_0$ so that $\mathcal{E}^{\text{rep}}$ and, hence, $\mathcal{E}$ increases with N^k.

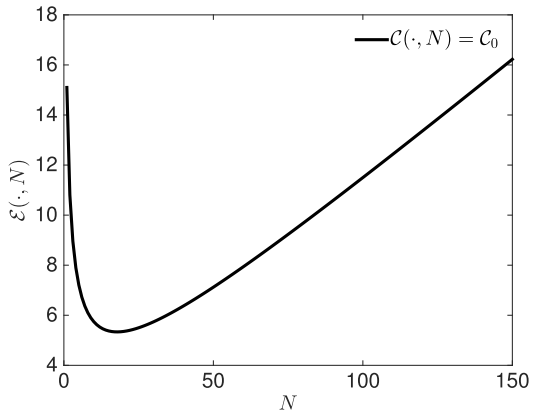

Figure 1. *Dependence of the simulation error $\mathcal{E}$ on the maximum number of affordable realizations N^k. The parameter values are set to $c_0 = 1, c_1 = 15, c_2 = 1, \mathcal{C}_0 = 10$ for illustration purposes.*

3.3. Resource-constrained model selection. Consider a pair of high- and low-fidelity models defined by the operators Ψ^{h} and Ψ^{l_i}, respectively. For the same computational cost $\mathcal{C}(\Psi^{\mathrm{h}}, N^{\mathrm{h}}) = \mathcal{C}(\Psi^{l_i}, N^{l_i}) = \mathcal{C}_0$ with $N^{\mathrm{h}} < N^{l_i}$, we define the relative error between the two models as $\Delta\mathcal{E} = \mathcal{E}(\Psi^{\mathrm{h}}, N^{\mathrm{h}}) - \mathcal{E}(\Psi^{l_i}, N^{l_i})$. It follows from (14) that

(15)
$$\Delta\mathcal{E} = \left[\frac{c_1\sqrt{c_0/\mathcal{C}_0}}{\left(\sqrt{|\Psi - \Psi^{\mathrm{h}}|} + \sqrt{|\Psi - \Psi^{l_i}|}\right)\sqrt{|\Psi - \Psi^{\mathrm{h}}||\Psi - \Psi^{l_i}|}} - c_2\right]\left(|\Psi - \Psi^{l_i}| - |\Psi - \Psi^{\mathrm{h}}|\right).$$

For the selected pair of models, the discrepancies $|\Psi - \Psi^{\mathrm{h}}|$ and $|\Psi - \Psi^{l_i}|$ are fixed and the relative error $\Delta\mathcal{E}$ is determined solely by the available computational resource $\mathcal{C}_0$, i.e., by the number of realizations of each model (N^{h} and N^{l_i}) one can afford. The relative error $\Delta\mathcal{E}$ decreases with the amount of allocated resource $\mathcal{C}_0$, changing its sign from positive to negative (Figure 2). The critical values of $\mathcal{C}_0$ and, thus, of N^{h} and N^{l_i} corresponding to $\Delta\mathcal{E} = 0$, are given by

(16)
$$\tilde{\mathcal{C}}_0 = \frac{c_0 c_1^2}{c_2^2|\Psi - \Psi^{l_i}||\Psi - \Psi^{\mathrm{h}}|\left(\sqrt{|\Psi - \Psi^{l_i}|} + \sqrt{|\Psi - \Psi^{\mathrm{h}}|}\right)^2},$$

(17)
$$\tilde{N}^{\mathrm{h}} = \frac{c_1^2}{c_2^2|\Psi - \Psi^{l_i}|\left(\sqrt{|\Psi - \Psi^{l_i}|} + \sqrt{|\Psi - \Psi^{\mathrm{h}}|}\right)^2},$$

(18)
$$\tilde{N}^{l_i} = \frac{c_1^2}{c_2^2|\Psi - \Psi^{\mathrm{h}}|\left(\sqrt{|\Psi - \Psi^{l_i}|} + \sqrt{|\Psi - \Psi^{\mathrm{h}}|}\right)^2}.$$

At this point, the model preference changes.

3.4. Resource-constrained optimal model. The fact that the simulation error $\mathcal{E}(\Psi^k, N)$ decreases with N, while the computational cost $\mathcal{C}(\Psi^k, N)$ increases, suggests the following

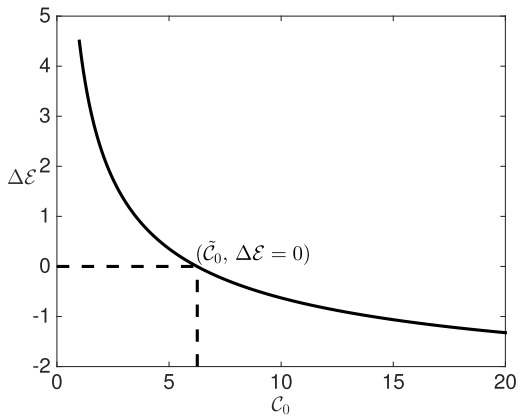

Figure 2. *Dependence of the relative error $\Delta\mathcal{E}$ on the allocated computational cost $\mathcal{C}_0$. The parameter values are set to $c_0 = 1, c_1 = 15, c_2 = 1, |\Psi - \Psi^h| = 1, |\Psi - \Psi^{l_i}| = 4$ for illustration purposes. The change in the sign of $\Delta\mathcal{E}$ indicates the reversal in the relative performance between the high- and low-fidelity models Ψ^h and Ψ^{l_i}.*

optimization problem. For a given computational cost $\mathcal{C}_0$ and a given set of the high- and low-fidelity models, minimize the simulation error $\mathcal{E}$. This gives rise to a cost function

$$\mathcal{L}\big(N, \Psi^k, \lambda\big) = \mathcal{E}(\Psi^k, N) + \lambda[\mathcal{C}(\Psi^k, N) - \mathcal{C}_0], \tag{19}$$

where λ is the Lagrange multiplier. For $\mathcal{C}$ and $\mathcal{E}$ in (12) and (10), setting to 0 the derivatives of $\mathcal{L}$ with respect to N, $|\Psi - \Psi^k|$, and λ yields the minimization point

$$\tilde{N} = \left(\frac{c_1\mathcal{C}_0}{2c_0c_2}\right)^{2/3}, \qquad |\Psi - \tilde{\Psi}| = \frac{c_0}{\mathcal{C}_0}\left(\frac{c_1\mathcal{C}_0}{2c_0c_2}\right)^{2/3}. \tag{20}$$

Substitution of (20) into (10) leads to the following proposition for the resource-constrained choice between high- and low-fidelity models.

Proposition 1. *Given a fixed computational resource $\mathcal{C}_0$ and the relationship between the cost $\mathcal{C}$, the sampling error $\mathcal{E}^{sam}$, and the model error $\mathcal{E}^{rep}$ in (12) and (10), the optimal model is such that*

$$\mathcal{E}^{sam} = \left(\frac{2c_0c_1^2c_2}{\mathcal{C}_0}\right)^{1/3}, \qquad \mathcal{E}^{rep} = \left(\frac{c_0c_1^2c_2}{4\mathcal{C}_0}\right)^{1/3}, \tag{21}$$

and the corresponding minimum error is

$$\mathcal{E}_{min} = \frac{(8c_0c_1^2c_2)^{1/3} + (c_0c_1^2c_2)^{1/3}}{(4\mathcal{C}_0)^{1/3}}. \tag{22}$$

Proposition 2. *Given a fixed computational resource $\mathcal{C}_0$ and the error model (14), the best model Ψ^b among a finite set of multifidelity models $\mathbf{\Psi} = \{\Psi^h, \Psi^{l_1}, \ldots, \Psi^{l_m}\}$ is defined as the one that minimizes the prediction error*

$$\Psi^b \triangleq \underset{\Psi^k \in \mathbf{\Psi}}{\operatorname{argmin}}\, \mathcal{E}(\Psi^k, \cdot), \qquad \mathcal{E}(\Psi^k, \cdot) = \sqrt{\frac{c_0}{\mathcal{C}_0}}\frac{c_1}{\sqrt{|\Psi - \Psi^k|}} + c_2|\Psi - \Psi^k| \tag{23}$$

and is subject to $\mathcal{E}(\Psi^b, \cdot) \geq \mathcal{E}(\tilde{\Psi}, \cdot)$.

4. EnKF. Similar to the accuracy of ensemble-based forward simulations with uncertain inputs analyzed in section 3, the accuracy of ensemble-based methods for data assimilation depends on the tradeoff between a model's fidelity and its computational cost. That is because system states' PDFs, or their statistics such as mean and variance, estimated from a small number of expensive high-fidelity model runs would have a higher sampling error $\mathcal{E}^{\text{sam}}$ (but smaller representation/model error $\mathcal{E}^{\text{rep}}$) than those estimated with a low-fidelity model. We investigate this tradeoff in the context of EnKF that is employed to assimilate a set of observations $\mathbf{y}_i = \{y_1, \ldots, y_i\}$ in (3) into the high- and low-fidelity models (2),

$$\hat{v}_{i+1} = \Psi^k(\hat{v}_i) + \xi_i^k, \quad k = \text{h and l}, \quad i \in \mathbb{N}; \qquad \hat{v}_0 \sim \mathcal{N}(\mu_0, C_0). \tag{24}$$

The sequential update of state v consists of two steps: prediction and analysis. At the former, one uses the Chapman–Kolmogorov equation,

$$f_{v_{i+1}|\mathbf{y}_i} = \int_{\mathbb{R}^N} f_{v_i|\mathbf{y}_i} f_{v_{i+1}|v_i} \mathrm{d}v_i, \tag{25}$$

to calculate the conditional PDF of v at time t_{i+1}, $f_{v_{i+1}|\mathbf{y}_i}$, from the conditional PDF of v at time t_i, $f_{v_i|\mathbf{y}_i}$. At the latter step, Bayes' theorem,

$$f_{v_{i+1}|\mathbf{y}_{i+1}} = \frac{f_{\mathbf{y}_{i+1}|v_{i+1}} f_{v_{i+1}|\mathbf{y}_i}}{f_{\mathbf{y}_{i+1}|\mathbf{y}_i}}, \tag{26}$$

is used to obtain the (posterior) PDF $f_{v_{i+1}|\mathbf{y}_{i+1}}$ from the (prior) PDF $f_{v_{i+1}|\mathbf{y}_i}$ when the measurement y_{i+1} is available.

Kalman filtering approximates the PDFs in (25) and (26) by their Gaussian counterparts, i.e., treats the random variables involved as

$$v_i|\mathbf{y}_i \sim \mathcal{N}(\mu_i, C_i), \quad v_{i+1}|\mathbf{y}_i \sim \mathcal{N}(\hat{\mu}_{i+1}, \hat{C}_{i+1}), \quad v_{i+1}|\mathbf{y}_{i+1} \sim \mathcal{N}(\mu_{i+1}, C_{i+1}). \tag{27}$$

This replaces the update step $f_{v_i|\mathbf{y}_i} \mapsto f_{v_{i+1}|\mathbf{y}_i}$ in (25) with $(\mu_i, C_i) \mapsto (\mu_{i+1}, C_{i+1})$, and the analysis step in (26) with

$$\exp\left(-\frac{1}{2}|v - \mu_{i+1}|^2_{C_{i+1}}\right) \propto \exp\left(-\frac{1}{2}|y_{i+1} - h(v)|^2_\Gamma - \frac{1}{2}|v - \hat{\mu}_{i+1}|^2_{\hat{C}_{i+1}}\right), \tag{28}$$

where $|\cdot|^2_A \equiv |\cdot|^2 A^{-1}$. To handle the nonlinearity of the high- and low-fidelity models in (24), we deploy an implementation based on the minimization principle [19]. Considering, for simplicity, a linear observation operator $h(v) = Hv$, this approach represents the update step as a quadratic minimization problem

$$\mu_{i+1} = \operatorname*{argmin}_v \Phi(v) \tag{29a}$$

with

$$\Phi(v) = \frac{1}{2}|y_{i+1} - Hv|^2_\Gamma + \frac{1}{2}|v - \hat{\mu}_{i+1}|^2_{\hat{C}_{i+1}}, \qquad \hat{\mu}_{i+1} = \Psi(\mu_i) + \xi_i. \tag{29b}$$

Algorithm 1. Implementation of the EnKF based on the minimization principle [19].

1. Initialization: at time step $i = 0$
   - Draw N samples, $v_0^{(n)}$ with $n = 1, \ldots, N$, of the initial state v_0 from the prior PDF, $v_0 \sim \mathcal{N}(\mu_0, C_0)$.
2. For $i = 1, 2, \ldots$
   (a) prediction (for $n = 1, \ldots, N$)
   - Use the high- or low-fidelity models in (24) to estimate the next state

$$\hat{v}_{i+1}^{(n)} = \Psi^k\left(\tilde{v}_i^{(n)}\right) + \xi_i^{k(n)}, \qquad n = 1, \ldots, N \tag{31}$$

   - Calculate the mean and covariance of the forecast

$$\hat{\mu}_{i+1} = \frac{1}{N}\sum_{n=1}^{N} \hat{v}_{i+1}^{(n)} \quad \text{and} \quad \hat{C}_{i+1} = \frac{1}{N-1}\sum_{n=1}^{N}\left(\hat{v}_{i+1}^{(n)} - \hat{\mu}_{i+1}\right)\left(\hat{v}_{i+1}^{(n)} - \hat{\mu}_{i+1}\right)^{\top} \tag{32}$$

   (b) analysis (for $n = 1, \ldots, N$)
   - Obtain a random data sample $y_{i+1}^{(n)}$ from the data model (3) with $\eta_{i+1} \sim \mathcal{N}(0, \Gamma)$

$$y_{i+1}^{(n)} = y_{i+1} + \eta_{i+1}^{(n)} \tag{33}$$

   - Calculate the Kalman gain in (30)

$$K_{i+1} = \hat{C}_{i+1} H^{\top}\left(H\hat{C}_{i+1}H^{\top} + \Gamma\right)^{-1} \tag{34}$$

   - Assimilate the model forecast $\hat{v}_{i+1}$ and data y_{i+1} to obtain N realizations of the "analyzed state" $\tilde{v}_{i+1}$

$$\tilde{v}_{i+1}^{(n)} = (I - K_{i+1}H)\hat{v}_{i+1}^{(n)} + K_{i+1}y_{i+1}^{(n)} \tag{35}$$

This minimization problem is solved by an update formula

$$\mu_{i+1} = (I - K_{i+1}H)\hat{\mu}_{i+1} + K_{i+1}y_{i+1}, \qquad K_{i+1} = \hat{C}_{i+1}H^T\left(H\hat{C}_{i+1}H^T + \Gamma\right)^{-1}. \tag{30}$$

Algorithm 1 implements the EnKF [7] by iteratively solving the above approximate Gaussian process and the update state.

The accuracy and stability of the EnKF have been investigated in [5, 2] for exact model operators. We analyze the accuracy of the EnKF for an imperfect model operator in section 4.1 and prove the assimilation asymptotic property at large times in section 4.2.

4.1. Accuracy of EnKF.

Proposition 3. *If the true system state $v(t)$ satisfies*

$$v_{i+1} = \Psi(v_i), \qquad v_0 = u, \tag{36}$$

and measurements of $v(t)$ at discrete times t_i are generated by adding bounded white noise to the true solution,

$$y_{i+1} = Hv_{i+1} + \eta_{i+1}, \quad \sup_{i\in\mathbb{N}} |\eta_{i+1}| = \eta < \infty, \tag{37}$$

then the estimation error of EnKF in Algorithm 1, $\mathcal{E}_{i+1} = \|\frac{1}{N}\sum_{n=1}^{N} \tilde{v}_{i+1}^{(n)} - v_{i+1}\|$, satisfies

$$\mathcal{E}_{i+1} \le \mathcal{E}_{i+1}^{sam} + \mathcal{E}_{i+1}^{rep} + \mathcal{E}_{i+1}^{ini} + \mathcal{E}_{i+1}^{dat}. \tag{38a}$$

The sampling ($\mathcal{E}_{i+1}^{sam}$), model ($\mathcal{E}_{i+1}^{rep}$), initialization ($\mathcal{E}_{i+1}^{ini}$), and data ($\mathcal{E}_{i+1}^{dat}$) errors at the $(i+1)$st time step are given by

$$\mathcal{E}_{i+1}^{sam} = \frac{1}{N}\left\| A_{i+1}\sum_{n=1}^{N}\left(\Psi^k\left(\tilde{v}_i^{(n)}\right) - \bar{v}_{i+1}\right)\right\| + \frac{1}{N}\left\|\sum_{n=1}^{N}\left[A_{i+1}\xi_i^{(n)} + K_{i+1}\eta_{i+1}^{(n)}\right]\right\|, \tag{38b}$$

$$\mathcal{E}_{i+1}^{rep} = \|A_{i+1}(\Psi^k(v_i) - \Psi(v_i))\|, \tag{38c}$$

$$\mathcal{E}_{i+1}^{ini} = \|A_{i+1}(\bar{v}_{i+1} - \Psi^k(v_i))\|, \tag{38d}$$

$$\mathcal{E}_{i+1}^{dat} = \|K_{i+1}\eta_{i+1}\|, \tag{38e}$$

where $A_{i+1} \equiv I - K_{i+1}H$ and $\bar{v}_{i+1} \equiv \mathbb{E}(\Psi^k(v_i|\boldsymbol{y}_i)) = \int_{\mathbb{R}^N} \Psi^k(v_i')\hat{f}_{\hat{v}_i|\boldsymbol{y}_i}(v_i')dv_i'$.

Proof. An estimate of the mean of $v(t_{i+1})$, computed from N realizations of either the exact model (36) or its high- or low-fidelity counterparts in (31), is

$$\mu_{i+1} = \frac{1}{N}\sum_{n=1}^{N}\tilde{v}_{i+1}^{(n)}, \tag{39}$$

Substituting (35) into this expression gives

$$\mu_{i+1} = A_{i+1}\frac{1}{N}\sum_{n=1}^{N}\hat{v}_{i+1}^{(n)} + K_{i+1}\frac{1}{N}\sum_{n=1}^{N}y_{i+1}^{(n)}. \tag{40}$$

For any model operator $\Psi^k \in \{\Psi, \Psi^{\mathrm{h}}, \Psi^{\mathrm{l}}\}$, accounting for (31), (33), and (37), this yields

$$\mu_{i+1} = A_{i+1}\frac{1}{N}\sum_{n=1}^{N}\left[\Psi^k\left(\tilde{v}_i^{(n)}\right) + \xi_i^{(n)}\right] + K_{i+1}\frac{1}{N}\sum_{n=1}^{N}\left[Hv_{i+1} + \eta_{i+1} + \eta_{i+1}^{(n)}\right]. \tag{41}$$

Let $\mathcal{E}_{i+1} = \|\mu_{i+1} - v_{i+1}\|$ define the solution error (in the ℓ_2 norm) at time step $i+1$. Recasting the true state (36) as

$$v_{i+1} = A_{i+1}\Psi(v_i) + K_{i+1}H\Psi(v_i) \tag{42}$$

and subtracting (42) from (41) yields

$$\mathcal{E}_{i+1} = \left\| A_{i+1}\left[\frac{1}{N}\sum_{n=1}^{N}\Psi^k\left(\tilde{v}_i^{(n)}\right) - \Psi(v_i)\right] + \frac{1}{N}\sum_{n=1}^{N}\left[A_{i+1}\xi_i^{(n)} + K_{i+1}\eta_{i+1}^{(n)}\right] + K_{i+1}\eta_i\right\|. \tag{43}$$

Let $\bar{v}_{i+1}$ denote the ensemble average corresponding to the sample mean $(1/N)\sum_{n=1}^{N}\Psi^k(\tilde{v}_i^{(n)})$. Then

$$\mathcal{E}_{i+1} = \left\| A_{i+1}\left[\frac{1}{N}\sum_{n=1}^{N}\left(\Psi^k\left(\tilde{v}_i^{(n)}\right)-\bar{v}_{i+1}\right)+\left(\bar{v}_{i+1}-\Psi^k(v_i)\right)+\left(\Psi^k(v_i)-\Psi(v_i)\right)\right]\right. \\ \left. +\frac{1}{N}\sum_{n=1}^{N}\left(A_{i+1}\xi_i^{(n)}+K_{i+1}\eta_{i+1}^{(n)}\right)+K_{i+1}\eta_i\right\|. \tag{44}$$

By virtue of the triangle inequality, this gives rise to (38). ∎

4.2. Asymptotic convergence of EnKF.

Theorem 4. *Suppose that the sampling ($\mathcal{E}_i^{sam}$), model ($\mathcal{E}_i^{rep}$), and data ($\mathcal{E}_i^{dat}$) errors in Proposition 3 remain bounded on the whole time horizon,*

$$\sup_{i\in\mathbb{N}}\mathcal{E}_i^{sam}=\mathcal{E}^{sam}<\infty,\qquad \sup_{i\in\mathbb{N}}\mathcal{E}_i^{rep}=\mathcal{E}^{rep}<\infty,\qquad \sup_{i\in\mathbb{N}}\mathcal{E}_i^{dat}=\mathcal{E}^{dat}<\infty. \tag{45}$$

Suppose also that $A_{i+1}\Psi^k : \mathbb{R}^{N_v}\to\mathbb{R}^{N_v}$ ($A_{i+1}\equiv I-K_{i+1}H$) is globally Lipschitz with constant $c<1$ in the ℓ_2 norm and that the impact of linearization $\mathbb{E}[\Psi(v)]\approx\mathbb{E}[\Psi(\mu)]$ is bounded, $\|A_{i+1}(\bar{v}_{i+1}-\Psi^k(\mu_i))\|\le\mathcal{E}^{avg}$. Then the convergence error of the assimilated result is bounded by

$$\limsup_{i\to\infty}\mathcal{E}_i\le\frac{\mathcal{E}^{sam}+\mathcal{E}^{rep}+\mathcal{E}^{dat}+\mathcal{E}^{avg}}{1-c}. \tag{46}$$

Proof. According to the law of large numbers, it follows from (38b) that

$$\mathcal{E}^{\mathrm{sam}}=\frac{1}{\sqrt{N}}\sup_{i\in\mathbb{N}}\left[A_{i+1}^2\left(\int_{\mathbb{R}^n}(\Psi(v_i')-\bar{v}_{i+1})^2 f_{v_i}(v_i')\mathrm{d}v_i'+\Sigma\right)+K_{i+1}^2\Gamma\right], \tag{47}$$

where Σ is the covariance of ξ_i. Since the Kalman gain K_{i+1} is determined by $\hat{C}_{i+1}$, so is $A_{i+1}\Psi^k$. The assumption of global Lipschitz continuity means

$$\|A_{i+1}(\Psi^k(\mu_i)-\Psi^k(v_i))\|\le c\,\mathcal{E}_i. \tag{48}$$

It follows from (48) and the linearization assumption that the initialization error satisfies

$$\mathcal{E}^{\mathrm{ini}}=\|A_{i+1}(\bar{v}_{i+1}-\Psi^k(\mu_i)+\Psi^k(\mu_i)-\Psi^k(v_i))\|\le\mathcal{E}^{\mathrm{avg}}+c\,\mathcal{E}_i. \tag{49}$$

Substituting (49) and (45) into (38) gives

$$\mathcal{E}_{i+1}\le c\mathcal{E}_i+\mathcal{E}^{\mathrm{sam}}+\mathcal{E}^{\mathrm{rep}}+\mathcal{E}^{\mathrm{dat}}+\mathcal{E}^{\mathrm{avg}}. \tag{50}$$

Application of the discrete time Gronwall lemma to (50) leads directly to (46). ∎

Corollary 5. *Consistency. If the EnKF in Algorithm 1 is used with the exact model (1) and if this model is linear, then Theorem 4 yields a bound*

$$\limsup_{i\to\infty, N\to\infty} \mathcal{E}_i \le \frac{\mathcal{E}^{dat}}{1-c}, \tag{51}$$

which is consistent with the Kalman filter result.

Proof. By definition, the representation error of the exact model Ψ is $\mathcal{E}^{\text{rep}} = 0$. If, in addition, the model is linear, then the linearization error is 0, which means $\mathcal{E}^{\text{avg}} = 0$. Finally, if the number of samples is infinite, then $\mathcal{E}^{\text{sam}} = 0$. Under these conditions, the bound in Theorem 4 reduces to (51).

Consider, next, the error of the standard Kalman filter. If μ_i' is the ensemble average of the state v_i, whose dynamics is governed by a linear operator $\Psi(\mu_i') = \psi\mu_i'$, then the mean of Kalman's analysis state is given by (30) and the true state by (42). Then,

$$\mathcal{E}_{i+1} = \|\mu_{i+1}' - v_{i+1}\| = \|(1 - K_{i+1}H)\psi(\mu_i' - v_i) + K_{i+1}(y_{i+1} - Hv_{i+1})\| \le c\mathcal{E}_i + \|K_{i+1}\eta\|.$$

The discrete time Gronwall lemma yields (51), which proves consistency. ∎

4.3. Impact of data on resource-constrained model selection. In the presence of data, the sampling error $\mathcal{E}^{\text{sam}}$ and the model error $\mathcal{E}^{\text{rep}}$, first introduced in subsection 3.1, are modified by the coefficient $I - K_{i+1}H$. Let $\alpha \sim I - K_{i+1}H$ and $\beta \sim K_{i+1}$ such that $\alpha + H\beta = I$. Then the error model (10) is replaced with

$$\mathcal{E}(\Psi^k, N|\mathbf{y}) = (\alpha c_1 + \beta c_3)/\sqrt{N} + \alpha c_2|\Psi - \Psi^k|, \tag{52}$$

where $c_3 \sim \Gamma$, and the errors $\mathcal{E}^{\text{ini}}$ and $\mathcal{E}^{\text{dat}}$ are omitted because they do not contribute to the model selection process. Since $\alpha \le I$, the availability of data always weakens the impact of model discrepancy, i.e., the effect of choosing a low-fidelity model. Furthermore, smaller measurement noise (variance) translates into smaller eigenvalues of α, which suggests both that better data have higher impact on model selection and that the model fidelity is less important when data quality is higher. These intuitive findings demonstrate the self-consistency of our analysis.

At every time step, the computational cost of EnKF in Algorithm 1 comprises the prediction and analysis steps. The cost of the prediction step equals the total runtime of N forward solves and is given by (12). The analysis step is executed when data are available; we assume that its runtime and cost are much smaller than those of a forward model solve. Furthermore, we assume the proportionality constant c_0 to be the same for all models; the case of c_0^k varying with the model fidelity (k = h and l) is treated in Appendix A. Neglecting the cost of the analysis step, the overall cost of EnKF is given by (12).

Theorem 6. *For the simulation error $\mathcal{E}$ in (52) and the simulation cost $\mathcal{C}$ in (12), and given a fixed computational cost $\mathcal{C}_0$ allocated for simulations, the optimal model and number of its realizations satisfy*

$$\tilde{N}_{\mathbf{y}} = \left(\frac{c_1\mathcal{C}_0}{2c_0c_2} + \frac{\beta}{\alpha}\frac{c_3\mathcal{C}_0}{2c_0c_2}\right)^{2/3}, \qquad |\Psi - \tilde{\Psi}_{\mathbf{y}}| = \frac{c_0}{\mathcal{C}_0}\left(\frac{c_1\mathcal{C}_0}{2c_0c_2} + \frac{\beta}{\alpha}\frac{c_3\mathcal{C}_0}{2c_0c_2}\right)^{2/3}. \tag{53}$$

Proof. In the presence of data, the cost function in (19) is modified as

$$\mathcal{L}(N, \Psi^k, \lambda) = \mathcal{E}(\Psi^k, N) + \lambda[\mathcal{C}(\Psi^k, N) - \mathcal{C}_0]. \tag{54}$$

For $\mathcal{E}$ and $\mathcal{C}$ in (52) and (12), setting to 0 the derivatives of $\mathcal{L}$ with respect to N, $|\Psi - \Psi^k|$, and λ yields the minimization point (53). ∎

The result in (54) has the following implication for the resource-constrained model selection in the presence of data. Since $\beta/\alpha \sim \hat{C}_{i+1} H^\top / \Gamma > 0$, both the optimized number of realizations and the allowed model discrepancy become larger. Hence, data availability argues in favor of selecting a lower-fidelity model that allows collecting a larger number of realizations during the allocated computing time. Furthermore, the higher the quality of data, the lower-fidelity model can be used. That is because the higher data quality means smaller variance Γ and bigger β/α. An example illustrating the impact of data quality on model selection is provided in Appendix B.

Proposition 7. *Given a fixed computational resource $\mathcal{C}_0$, the best model among a finite set of multifidelity models $\mathbf{\Psi} = \{\Psi^h, \Psi^{l_1}, \ldots, \Psi^{l_m}\}$ for assimilating sequential observational data by EnKF is such that*

$$\Psi^b_{\mathbf{y}} \triangleq \operatorname*{argmin}_{\Psi^k \in \mathbf{\Psi}} \mathcal{E}(\Psi^k, \cdot|\mathbf{y}), \qquad \mathcal{E}(\Psi^k, \cdot|\mathbf{y}) = \sqrt{\frac{c_0}{\mathcal{C}_0}} \frac{\alpha c_1 + \beta c_3}{\sqrt{|\Psi - \Psi^k|}} + \alpha c_2 |\Psi - \Psi^k| \tag{55}$$

and is subject to $\mathcal{E}(\Psi^b_{\mathbf{y}}, \cdot|\mathbf{y}) \geq \mathcal{E}(\tilde{\Psi}_{\mathbf{y}}, \cdot|\mathbf{y})$.

5. Computational examples.

We provide two examples that illustrate the applicability of our model selection method. The first example deals with an ordinary differential equation, for which a sequence of lower-fidelity models is constructed by progressively increasing the time step used in its discretization. The second example comprises the viscous Burgers equation as the high-fidelity model and a linear advection-diffusion equation as its low-fidelity counterpart. In both cases, the initial conditions of high- and low-fidelity models are deterministic but unknown.

5.1. Ordinary differential equation.

Consider a system state $v(t) : \mathbb{R}^+ \to \mathbb{R}^+$ whose true dynamics is governed by an ordinary differential equation

$$\frac{\mathrm{d}v}{\mathrm{d}t} = \alpha v, \qquad v(t=0) = u, \qquad t \in [0, T], \tag{56}$$

with the deterministic constant $\alpha \in \mathbb{R}^+$ and initial state $u \in \mathbb{R}^+$. Its discretized true solution is

$$v_{i+1} = v_i \, \mathrm{e}^{-\alpha \Delta t}, \qquad i \geq 0, \qquad v_0 = u, \tag{57}$$

where Δt denotes the discretization time step. Noisy measurements of the state $v(t)$ are generated with a model

$$y_{i+1} = v_{i+1} + \epsilon_i, \tag{58}$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma^2_\epsilon)$ are i.i.d. Gaussian variables.

A sequence of multifidelity models is generated by solving (56) with the Euler method,

$$\hat{v}_{i+1}^k = \hat{v}_i(1+\alpha\Delta t) + \xi_i^k, \qquad v_0 \sim \mathcal{N}(\mu_0, \sigma_0^2), \tag{59}$$

where $\xi_i^k \sim \mathcal{N}(0, \sigma_\xi^2)$ are i.i.d. Gaussian variables. The high-fidelity model is obtained by using a time step Δt_{h}, and low-fidelity models are built by using larger time steps Δt_{l_i} $(i = 1, 2, \ldots)$ such that $\Delta t_{\mathrm{h}} < \Delta t_{l_1} < \Delta t_{l_2} < \cdots$. The observation equation is described as

$$y_{i+1} = Hv_{i+1} + \eta_{i+1}, \tag{60}$$

where $\eta_i \sim \mathcal{N}(0, \sigma_\eta^2)$ are i.i.d. Gaussian variables. In the simulations reported below, we set $\alpha = 1.5$, $u = 1$, $\mu_0 = 0$, $\sigma_{v_0}^2 = 1$, $\sigma_\eta^2 = 1$, and $H = 1$. For simplicity, neither model error nor data generating error is considered here, i.e., $\sigma_\xi^2 = 0$ and $\sigma_\epsilon^2 = 0$. The sequence of time steps is $\Delta t_k = \{0.001, 0.002, 0.005, 0.01, 0.015, 0.03\}$, with the first number corresponding to the high-fidelity model and the rest forming a set of the low-fidelity models.

Given the time step Δt_k, EnKF relies on $N = \{5, 10, 20, 50, 100, 200, 500, 1000, 2000\}$ realizations. The simulation is conducted up to time $T = 3$, with the measurements available at times $t = 0.09k$, where $k = 1, 2, \ldots, 33$. In the absence of data, (59) is used to propagate the analyzed states without assimilation. Every data assimilation experiment is repeated 30 times, and the results are averaged in order to reduce the error arising from pseudorandom number generator sampling.

We compute the cost $\mathcal{C}$ as the runtime T_{run} of our EnKF algorithm for different time steps Δt and sampling numbers N. The error $\mathcal{E}$ is the average of the difference $\bar{E}$ between the true value and its assimilated counterpart in the ℓ_1 norm over the whole time domain, i.e.,

$$\bar{E} = \frac{1}{N_t}\sum_i^{N_t} |v_i - \tilde{\mu}_i|, \qquad N_t = \frac{T}{\Delta t}, \qquad \tilde{\mu}_i = \frac{1}{N}\sum_n^N \tilde{v}_i^{(n)}. \tag{61}$$

For the problem under consideration, the discrepancy between the models (57) and (59) is $|\Psi - \Psi^i| = |\exp(\alpha\Delta t) - (1+\alpha\Delta t)|$, and our general definitions of the cost function $\mathcal{C}(\Psi, N)$ in (12) and the error $\mathcal{E}(\Psi, N)$ in (10) take the form

$$\mathcal{C}(\Delta t, N) = \frac{c_0 N}{|\mathrm{e}^{\alpha\Delta t} - (1+\alpha\Delta t)|}, \qquad \mathcal{E}(\Delta t, N) = \frac{c_1}{N^{\alpha_1}} + c_2|\mathrm{e}^{\alpha\Delta t} - (1+\alpha\Delta t)|. \tag{62}$$

The simulation results reported in Figure 3 reveal these definitions to be accurate, with the exponent $\alpha_1 = 1$ (rather than $\alpha_1 = 1/2$ used in (10) for illustration purposes) and the proportionality constants $c_0 = 8 \cdot 10^{-7}$, $c_1 = 1$, and $c_2 = 2.5$.

Figure 4 exhibits the observed dependence of the simulation error $\mathcal{E}$ on the number of realizations N for the fixed cost $\mathcal{C} = \mathcal{C}_0 = 1000c_0$. To identify the optimal model at such cost, we construct a set of time steps $\mathrm{U}_{\Delta t} = \{0.001, 0.002, 0.004, 0.005, 0.010, 0.020, 0.030)\}$. The simulation cost $\mathcal{C}$, i.e., the runtime T_{run}, is approximately the product of the total number of time steps N_t and the number of realizations N, i.e., $\mathcal{C} = NT/\Delta t$. For $\mathcal{C}_0 = 9000$ and $T = 3$, this yields a set of Monte Carlo realizations $\mathrm{U}_N = 3000U_{\Delta t} = \{3, 6, 12, 15, 30, 60, 90\}$. Note that the latter is nearly identical to $\mathrm{U}'_N = 1000|\mathrm{e}^{-\alpha\Delta t} - (1-\alpha\Delta t)| = \{3, 6, 12, 15, 30, 60, 91\}$ predicted by the theory (62). This indicates the correctness of our definition (62).

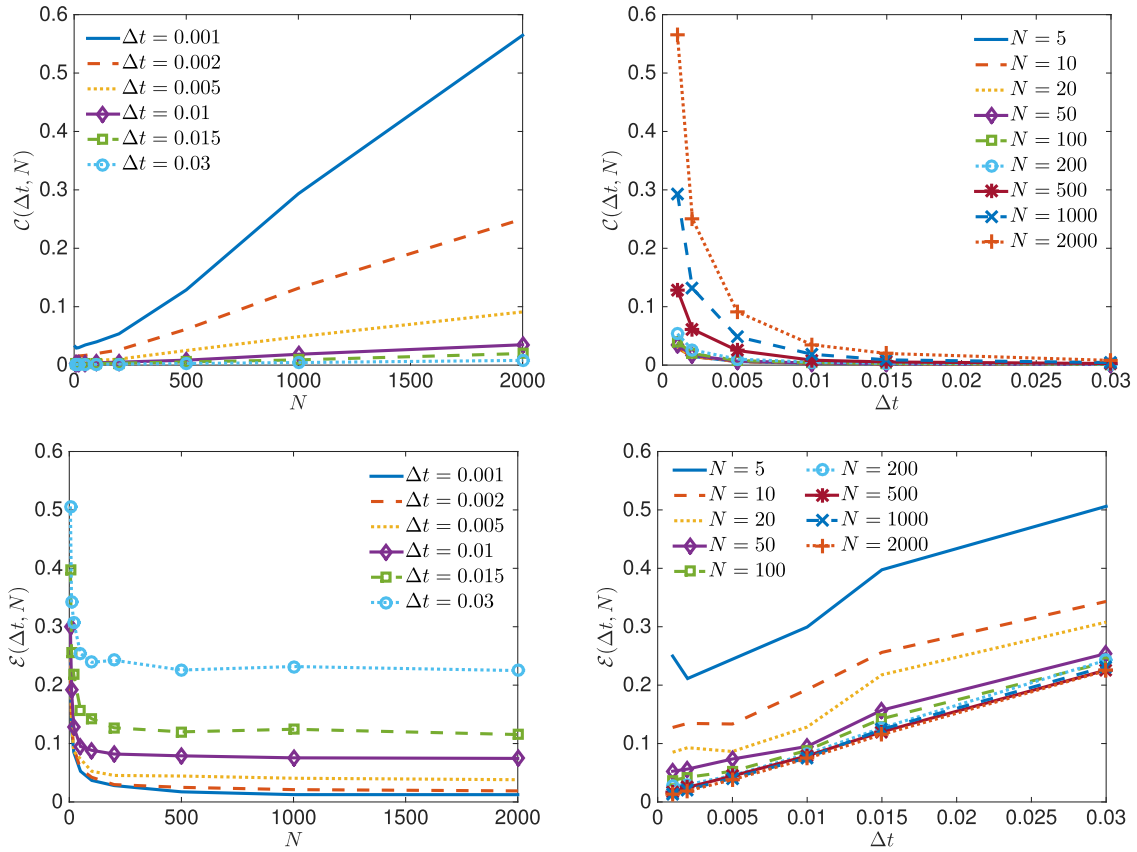

Figure 3. *Dependence of the simulation cost $\mathcal{C} = T_{run}$ (top row) and simulation error $\mathcal{E} = \bar{E}$ on the number of Monte Carlo realizations N (left column) and the time step Δt (right column).*

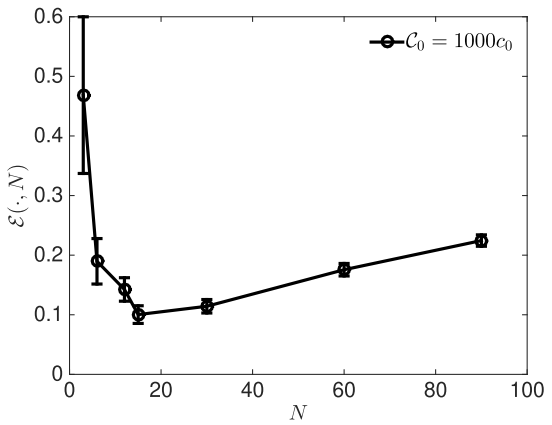

Figure 4. *Dependence of the simulation error $\mathcal{E}$ on the number of realizations N of multifidelity models for the given computational cost $\mathcal{C}_0 = 1000c_0$. The error bars represent the 95% confidence interval computed from 30 repeated simulations.*

As predicted by the general considerations reported in Figure 1, there is an optimal low-fidelity model (corresponding to the time step Δt_{l_k}) that has the lowest simulation error $\mathcal{E}$. This l_kth model is identified by the number N^{l_k} that minimizes $\mathcal{E}(N)$. For the simulations reported in Figure 4, this number is $N^{l_k} = 15$, which corresponds to the low-fidelity model with $\Delta t_{l_k} = 0.005$.

5.2. Burgers' equation. Consider, next, the nonlinear viscous Burgers equation

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = \nu\frac{\partial^2 u}{\partial x^2}, \qquad |x| \le 1, \quad t > 0, \tag{63}$$

with the viscosity coefficient $\nu \in \mathbb{R}^+$ and initial and boundary conditions

$$u(x,0) = -\sin(\pi x), \qquad u(1,t) = u(-1,t) = 0. \tag{64}$$

This problem admits an exact analytical solution [1],

$$u(x,t) = \frac{\displaystyle\int_{-\infty}^{\infty} \sin[\pi(x-\eta)]g(x-\eta)\exp[-\eta^2/(4\nu t)]\mathrm{d}\eta}{\displaystyle\int_{-\infty}^{\infty} g(x-\eta)\exp[-\eta^2/(4\nu t)]\mathrm{d}\eta}, \qquad g(x) = \exp\left(-\cos\frac{\pi x}{2\pi\nu}\right), \tag{65}$$

obtained by means of the Cole–Hopf transformation. The "true solution" is obtained by using the Hermite quadrature to evaluate the integrals in (65).

High- and low-fidelity approximations of (65) have an uncertain initial state $u(x,0) = v$. We model this state as a Gaussian random variable, $v \sim \mathcal{N}(0,1)$. To obtain a high-fidelity model, we rewrite (63) in the conservative form,

$$\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = \nu\frac{\partial^2 u}{\partial x^2}, \qquad f(u) = \frac{u^2}{2}, \tag{66}$$

and approximate it by using an explicit scheme in time and the central-difference scheme in space,

$$u_j^{n+1} = u_j^n - \frac{\Delta t}{2\Delta x}\left[f\left(u_{j+1}^n\right) - f\left(u_{j-1}^n\right)\right] + \frac{\nu\Delta t}{\Delta x^2}\left(u_{j+1}^n - 2u_j^n + u_{j-1}^n\right) + \xi^n, \tag{67}$$

where $u_j^n \equiv u(x_j, t_n)$ and the random noise ξ^n represents the temporally fluctuating model error.

A low-fidelity model is a linearized version of the Burgers equation (63),

$$\frac{\partial u}{\partial t} + u^*\frac{\partial u}{\partial x} = \nu\frac{\partial^2 u}{\partial x^2}, \qquad |x| \le 1, \quad t > 0, \tag{68}$$

subject to the same initial and boundary conditions (64). The constant advection velocity u^* is set to u_{avg}. The same numerical discretization yields

$$u_j^{n+1} = u_j^n - u^*\frac{\Delta t}{\Delta x}\left(u_{j+1}^n - u_{j-1}^n\right) + \frac{\nu\Delta t}{\Delta x^2}\left(u_{j+1}^n - 2u_j^n + u_{j-1}^n\right) + \xi^n. \tag{69}$$

If the grid functions are organized as vectors, then (69) takes the form

$$\mathbf{u}^{n+1} = \mathbf{Q}\mathbf{u}^n + \boldsymbol{\xi}^n, \tag{70}$$

where $\mathbf{u}^n = (u_1^n, u_2^n, \dots, u_p^n)^\top$, $\boldsymbol{\xi}^n = (\xi_1^n, \xi_2^n, ..., \xi_p^n)^\top \sim \mathcal{N}(0, \boldsymbol{\Sigma})$ is an i.i.d. Gaussian vector with components $\Sigma_{ij} = \Sigma\delta_{ij}$, and

$$\mathbf{Q} = \begin{pmatrix} 1 - \dfrac{2\nu\Delta t}{\Delta x^2} & -\dfrac{u^*\Delta t}{2\Delta x} + \dfrac{\nu\Delta t}{\Delta x^2} & \dots & 0 & \dfrac{u^*\Delta t}{2\Delta x} + \dfrac{\nu\Delta t}{\Delta x^2} \\ \dfrac{u^*\Delta t}{2\Delta x} + \dfrac{\nu\Delta t}{\Delta x^2} & 1 - \dfrac{2\nu\Delta t}{\Delta x^2} & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -\dfrac{u^*\Delta t}{2\Delta x} + \dfrac{\nu\Delta t}{\Delta x^2} & 0 & \dots & \dfrac{u^*\Delta t}{2\Delta x} + \dfrac{\nu\Delta t}{\Delta x^2} & 1 - \dfrac{2\nu\Delta t}{\Delta x^2} \end{pmatrix}_{p\times p}.$$

The linear observation operator $\mathbf{H}$ and observation noise $\boldsymbol{\eta}$ are used to generate data $\mathbf{y}$ in accordance with

$$\mathbf{y}^{n+1} = \mathbf{H}\mathbf{u}^{n+1} + \boldsymbol{\eta}^n. \tag{71}$$

Here, $\mathbf{y}^n = (y_1^n, y_2^n, \dots, y_m^n)^\top$, $\boldsymbol{\eta}^n = (\eta_1^n, \eta_2^n, \dots, \eta_m^n)^\top \sim \mathcal{N}(0, \boldsymbol{\Gamma})$ is an i.i.d. Gaussian vector with components $\Gamma_{ij} = \Gamma\delta_{ij}$, and

$$\mathbf{H} = \begin{pmatrix} 1 & 0 & \dots & 0 & 0 & \dots & 0 & 0 & \dots \\ 0 & 0 & \dots & 1 & 0 & \dots & 0 & 0 & \dots \\ \vdots & \vdots & & \vdots & \vdots & \ddots & \vdots & \vdots & \\ 0 & 0 & \dots & 0 & 0 & \dots & 1 & 0 & \dots \end{pmatrix}_{m\times p}.$$

In the simulations reported below, the multifidelity models (67) and (70) are solved on the space-time domain $x \times t \in [-1, 1] \times [0, 3/\pi]$, discretized with $\Delta x = 2/121$ and $\Delta t = 0.03/\pi$. The parameters are set to $u^* = u_{\text{avg}} = 0$, $\nu = 0.01/\pi$, $\Sigma = 0.01$, and $\Gamma = 0.05$. The number of observation points is $m = 20$; these are equally spaced on the interval $[-1, 1]$ and are available at every time step. We define the cost function $\mathcal{C}$ as the runtime of the forecast equation solved $N = \{20, 35, 50, 75, 100, 200, 350, 500, 750, 1000, 1500, 2000, 5000\}$ times (realizations), and the simulation error $\mathcal{E}$ as the spatial average of the distance (under ℓ_1 norm) between the assimilated result and the truth at the last time step. This numerical experiment is repeated 30 times, and the results are averaged in order to reduce the error arising from pseudorandom number generator sampling.

Figure 5 reveals that when the number of realizations $\ln N \gtrsim 8$, the error of the high-fidelity model (67) is smaller than that of its low-fidelity counterpart (70) at the same computational cost. This result is to be expected, since the model fidelity dominates the simulation error when the number of samples is sufficiently large, in accordance with Theorem 4. On the other hand, when the number of realizations is relatively small ($\ln N \lesssim 3$), the error of high-fidelity model (67) exceeds that of its low-fidelity counterpart (70) at the same computational cost. Therefore, as the fixed computational cost $\mathcal{C}_0$ increases, first the low-fidelity model outperforms its high-fidelity counterpart, and then the relative performance of these two models switches. This is consistent with the general analysis in Figure 2.

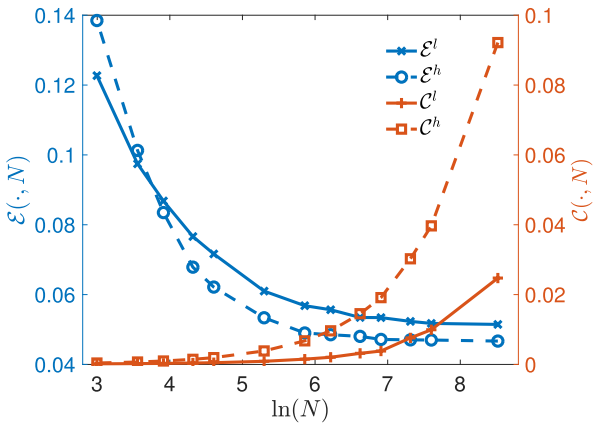

Figure 5. *Dependence of the simulation error $\mathcal{E}$ (left axis and blue curves) and simulation cost $\mathcal{C}$ (right axis and red curves) on the number of realizations of the multifidelity models (67) and (70). The dashed and solid lines represent the high- and low-fidelity models, respectively.*

6. Summary. We developed a general framework for analysis of the impact of data assimilation on cost-constrained model selection. The framework relies on the definitions of cost and accuracy functions in the context of data assimilation for multifidelity models with uncertain (random) coefficients, and contains an estimate of error bounds for a system's state prediction obtained by assimilating data into a model via EnKF. This estimate is given in terms of model error, sampling error, and data error. We provided two examples that illustrate the applicability of our model selection method. The first example deals with an ordinary differential equation, for which a sequence of lower-fidelity models is constructed by progressively increasing the time step used in its discretization. The second example comprises the viscous Burgers equation as the high-fidelity model and a linear advection-diffusion equation as its low-fidelity counterpart.

Our analysis leads to the following major conclusions.

- Our definitions of the computational cost ($\mathcal{C}$) and accuracy (expressed in terms of corresponding error $\mathcal{E}$) of sampling-based (e.g., Monte Carlo) solutions of multifidelity models require an assumption on the functional dependence of $\mathcal{C}$ and $\mathcal{E}$ on the number of Monte Carlo realizations, N.
- The two examples considered confirm the validity of the assumed functional forms $\mathcal{C} = \mathcal{C}(N)$ and $\mathcal{E} = \mathcal{E}(N)$.
- When N is small, the sampling error dominates the simulation error $\mathcal{E}$, and, hence, the lower-fidelity model gives more accurate predictions than its higher-fidelity counterpart.
- As N becomes sufficiently large, the model error dominates the simulation error $\mathcal{E}$, which argues for the use of the higher-fidelity model.
- In the cost-constrained model selection, the computation cost $\mathcal{C}^*$ is fixed so that the number of Monte Carlo realizations is determined by inverting the function $\mathcal{C} = \mathcal{C}(N)$.
- The availability of data, assimilated by means of EnKF, always weakens the impact of model discrepancy, i.e., the effect of choosing a low-fidelity model.

- Data availability argues in favor of selecting a lower-fidelity model that allows collecting a larger number of realizations during the allocated computing time. The higher the quality of data, the lower-fidelity model can be used.

Several questions remain open and will be investigated in follow-up studies. These include the impact of data assimilation techniques other than EnKF (e.g., particle filter and smoothing problem) on the cost-accuracy tradeoff for multifidelity models, and integration of data assimilation methods into multifidelity simulations (e.g., in the context of multilevel Monte Carlo).

Appendix A. General cost and accuracy functions.

In general, the constants of proportionality c_0^k, c_1^k, and c_2^k in the error models (8) and (12) vary between multifidelity models, i.e., with k. This dependence stems from multiple sources, including the model type denoted by a categorical variable s_k (e.g., $s_k \in \{$interpolation, regression, simplified-physics, projection$\}$); the number of parameters, N_{par}, that quantifies model complexity; and the number of degrees of freedom in a model, N_{deg}, as quantified by, e.g., a mesh size Δ_k (small Δ_k results in large N_{deg} and vice versa). Hence, $c_i^k = c_i(s_k, N_{\text{par}}^k, N_{\text{deg}}^k)$ for $i = 1, 2, 3$ and $k = \text{h, l}$.

As an example, we consider a set of models that differ only in their degrees of complexity (N_{par}^k), i.e., come from the same class s and have the same number of degrees of freedom N_{deg}. A higher model complexity has been found to reduce the model's generality and, hence, its prediction accuracy [12]. (As a caveat, we note that this finding is based on polynomial approximations and seems to contradict recent studies that use "deep learning" or neural-network approximations.) In our context, this observation is codified in the monotonically increasing dependence $c_2^k = c_2(N_{\text{par}}^k)$. It is also reasonable to surmise that model complexity affects the simulation cost, so that $c_0^k = c_0(N_{\text{par}}^k)$ is an increasing function. The remaining proportionality constant is kept unchanged between the models, $c_1^k = c_1$, because model complexity is not expected to affect the sampling error $\mathcal{E}^{\text{sam}}$. With these assumptions about model complexity, the cost and accuracy functions (12) and (10) take the form

(72)
$$\mathcal{C}(\Psi^k, N) = c_0\left(N_{\text{par}}^k\right)\frac{N}{|\Psi - \Psi^k|}, \qquad \mathcal{E}(\Psi^k, N) = \frac{c_1}{\sqrt{N}} + c_2\left(N_{\text{par}}^k\right)|\Psi - \Psi^k|, \qquad k = \text{h, l}.$$

A higher model fidelity typically requires a larger N_{par}^k and a smaller $|\Psi - \Psi^k|$. Hence, the term $c_2(N_{\text{par}}^k)|\Psi - \Psi^k|$ suggests that an optimal prediction accuracy, smallest $\mathcal{E}(\Psi^k, N)$, is obtained by setting a reasonable N_{par}^k, which is consistent with previous findings [12]. Furthermore, the simulation error $\mathcal{E}$ depends on model fidelity and the simulation cost $\mathcal{C}_0$ as

(73)
$$\mathcal{E}(\Psi^k, \mathcal{C}_0) = \sqrt{\frac{c_0\left(N_{\text{par}}^k\right)}{\mathcal{C}_0}}\frac{c_1}{\sqrt{|\Psi - \Psi^k|}} + c_2\left(N_{\text{par}}^k\right)|\Psi - \Psi^k|, \qquad k = \text{h, l}.$$

With the parameters c_0 and c_2 depending on model complexity, we use the results of section 3 to select a model. According to Proposition 2, the best model is such that

(74)
$$\Psi^b \triangleq \underset{\Psi^k}{\operatorname{argmin}}\, \mathcal{E}(\Psi^k, \mathcal{C}_0), \qquad k = \text{h, l}.$$

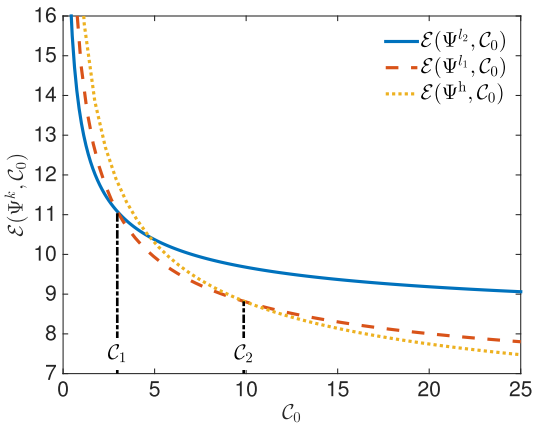

Figure 6. *Dependence of the simulation error $\mathcal{E}$ on the simulation cost $\mathcal{C}_0$ with different model complexity $N_{par}^h = 3$, $N_{par}^{l_1} = 2$, $N_{par}^{l_2} = 1$. The proportionality constants are set to $c_1 = 15$, $c_0^k = c_2^k = \sqrt{N_{par}^k}$ and $|\Psi^{l_2} - \Psi| = 8$, $|\Psi^{l_1} - \Psi| = 4.3$, $|\Psi^h - \Psi| = 3$ for illustration purposes.*

Figure 6 shows this dependence for $c_0 = c_2 = (N_{\mathrm{par}}^k)^{1/2}$ under different allocations of simulation cost. The values $\mathcal{C}_1$ and $\mathcal{C}_2$ are turning points for the process of model selection: Ψ^{l_2} has the lowest prediction error when $\mathcal{C}_0 \leq \mathcal{C}_1$; as the allocated cost increases, Ψ^{l_1} is the optimal model when $\mathcal{C}_1 < \mathcal{C}_0 < \mathcal{C}_2$, and Ψ^{h} is the best candidate when $\mathcal{C}_0 \geq \mathcal{C}_2$.

Appendix B. Impact of data quality on model selection.

Suppose that at a certain time t_i, the system is sampled M times giving a data set $\{y_{i,1}, \ldots, y_{i,M}\}$. The data are unbiased and, accounting for measurement error, treated as i.i.d. Gaussian random variables, $y_{i,m} \sim \mathcal{N}(\bar{y}_i, \Gamma)$, where $\bar{y}_i = h(v_i)$ in accordance with (3). Then the sample mean and variance obey the strong law of large numbers:

$$\hat{y}_i = \frac{1}{M}\sum_{m=1}^{M} y_{i,m}, \qquad \sigma_{y_i}^2 = \frac{\Gamma}{M}, \qquad \mathbb{P}\left[\lim_{M\to\infty} \hat{y}_i = h(v_i)\right] = 1, \tag{75}$$

i.e., the quality of data $\hat{y}_i$ increases with M, so that the measurements provide very precise information about the true states at time t_i. In the analysis step in Algorithm 1, we replace the observation y_{i+1} with the average $\hat{y}_{i+1}$; then the random data sample (33) is expressed as

$$y_{i+1}^{(n)} = \hat{y}_{i+1} + \hat{\eta}_{i+1}^{(n)} \quad \text{where} \quad \hat{\eta}_{i+1} \sim \mathcal{N}(0, \Gamma/M). \tag{76}$$

The Kalman gain in (34) turns into

$$K_{i+1} = \hat{C}_{i+1} H^\top \left(H\hat{C}_{i+1}H^\top + \Gamma/M\right)^{-1}, \tag{77}$$

which yields

$$I - K_{i+1}H = \frac{\Gamma}{M}\left(H\hat{C}_{i+1}H^\top + \Gamma/M\right)^{-1}. \tag{78}$$

As the number of measurements M increases, $\beta/\alpha \sim M\hat{C}_{i+1}H^\top/\Gamma$ increases. According to (53), larger values of M and, hence, higher data quality weaken the impact of model fidelity on the model selection.

REFERENCES

[1] C. Basdevant, M. Deville, P. Haldenwang, J. M. Lacroix, J. Ouazzani, R. Peyret, P. Orlandi, and A. T. Patera, *Spectral and finite difference solutions of the Burgers equation*, Comput. & Fluids, 14 (1986), pp. 23–41.

[2] C. E. A. Brett, K. F. Lam, K. J. H. Law, D. S. McCormick, M. R. Scott, and A. M. Stuart, *Accuracy and stability of filters for dissipative PDEs*, Phys. D, 245 (2013), pp. 34–45.

[3] S. T. Buckland, K. P. Burnham, and N. H. Augustin, *Model selection: An integral part of inference*, Biometrics, 53 (1997), pp. 603–618.

[4] K. P. Burnham and D. R. Anderson, *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, Springer Science & Business Media, New York, 2003.

[5] A. Carrassi, M. Ghil, A. Trevisan, and F. Uboldi, *Data assimilation as a nonlinear dynamical systems problem: Stability and convergence of the prediction-assimilation system*, Chaos, 18 (2008), 023112.

[6] G. Claeskens and N. L. Hjort, *Model Selection and Model Averaging*, Cambridge University Press, Cambridge, UK, 2008.

[7] G. Evensen, *The ensemble Kalman filter: Theoretical formulation and practical implementation*, Ocean Dyn., 53 (2003), pp. 343–367.

[8] M. B. Giles, *Multilevel Monte Carlo methods*, Acta Numer., 24 (2015), pp. 259–328.

[9] H. Hoel, K. J. H. Law, and R. Tempone, *Multilevel ensemble Kalman filtering*, SIAM J. Num. Anal., 54 (2016), pp. 1813–1839.

[10] J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky, *Bayesian model averaging: A tutorial*, Stat. Sci., (1999), pp. 382–401.

[11] G. Lin, A. M. Tartakovsky, and D. M. Tartakovsky, *Uncertainty quantification via random domain decomposition and probabilistic collocation on sparse grids*, J. Comput. Phys., 229 (2010), pp. 6995–7012.

[12] I. J. Myung, *The importance of complexity in model selection*, J. Math. Psych., 44 (2000), pp. 190–204.

[13] A. Narayan, C. Gittelson, and D. Xiu, *A stochastic collocation algorithm with multifidelity models*, SIAM J. Sci. Comput., 36 (2014), pp. A495–A521.

[14] A. Narayan, Y. Marzouk, and D. Xiu, *Sequential data assimilation with multiple models*, J. Comput. Phys., 231 (2012), pp. 6401–6418.

[15] B. Peherstorfer, K. Willcox, and M. Gunzburger, *Optimal model management for multifidelity Monte Carlo estimation*, SIAM J. Sci. Comput., 38 (2016), pp. A3163–A3194.

[16] B. Peherstorfer, K. Willcox, and M. Gunzburger, *Survey of multifidelity methods in uncertainty propagation, inference, and optimization*, SIAM Rev., 60 (2018), pp. 550–591.

[17] M. Sinsbeck and D. M. Tartakovsky, *Impact of data assimilation on cost-accuracy tradeoff in multifidelity models*, SIAM/ASA J. Uncertain Quantif., 3 (2015), pp. 954–968.

[18] I. M. Sobol, *A Primer for the Monte Carlo Method*, CRC Press, Boca Raton, FL, 2018.

[19] A. Stuart and K. Zygalakis, *Data Assimilation: A Mathematical Introduction*, Tech. report, Oak Ridge National Laboratory, Oak Ridge, TN, 2015.

[20] S. Taverniers and D. M. Tartakovsky, *Estimation of distributions via multilevel Monte Carlo with stratified sampling*, J. Comput. Phys., 419 (2020), 109572.

[21] L. Yang, A. Narayan, and P. Wang, *Sequential data assimilation with multiple nonlinear models and applications to subsurface flow*, J. Comput. Phys., 346 (2017), pp. 356–368.