Massachusetts Institute of Technology

Output error behavior for discretizations of ergodic, chaotic systems of ordinary differential equations 🅕

🆔 **Cory V. Frontin** and 🆔 **David L. Darmofal**

COLLECTIONS

Ⓕ This paper was selected as Featured

View Online   Export Citation   CrossMark

ARTICLES YOU MAY BE INTERESTED IN

Flow and acoustics of unmanned vehicles
Physics of Fluids **34**, 100402 (2022); https://doi.org/10.1063/5.0129577

Deep reinforcement learning for large-eddy simulation modeling in wall-bounded turbulence
Physics of Fluids **34**, 105132 (2022); https://doi.org/10.1063/5.0106940

On Oreology, the fracture and flow of "milk's favorite cookie$^{®}$"
Physics of Fluids **34**, 043107 (2022); https://doi.org/10.1063/5.0085362
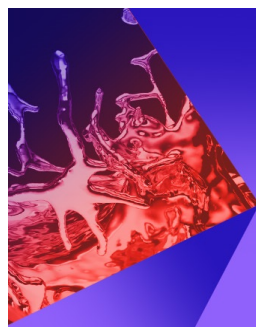
Physics of Fluids

Special Topic: Paint and Coating Physics

Submit Today!

AIP Publishing

Output error behavior for discretizations of ergodic, chaotic systems of ordinary differential equations

View Online   Export Citation   CrossMark

Cory V. Frontin[a] (iD) and David L. Darmofal[b] (iD)

AFFILIATIONS

Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

[a] Author to whom correspondence should be addressed: cfrontin@mit.edu
[b] Electronic mail: darmofal@mit.edu

ABSTRACT

The use of numerical simulation for prediction of characteristics of chaotic dynamical systems inherently involves unpredictable processes. In this work, we develop a model for the expected error in the simulation of ergodic, chaotic ordinary differential equation (ODE) systems, which allows for discretization and statistical effects due to unpredictability. Using this model, we then generate a framework for understanding the relationship between the sampling cost of a simulation and the expected error in the result and explore the implications of the various parameters of simulations. Finally, we generalize the framework to consider the total cost—including unsampled spin-up timesteps—of simulations and consider the implications of parallel computational environments to give a realistic model of the relationship between wall-clock time and the expected error in simulation of a chaotic ODE system.

I. INTRODUCTION

Progress in computational processing capabilities and numerical methods has led to increasing use of unsteady, scale resolving simulation methods for turbulent flows for applied research.[1–3] In spite of advances, the cost of these simulations, particularly large-eddy simulation (LES), remains an obstacle to industrial use. Cost reduction remains a key research area for these systems, including more aggressive modeling approaches.[1,4] On the other hand, advancements in numerical methods have demonstrated that a stable, high-order solution of turbulent systems is possible[5] with ongoing developments enabling new efficiencies.[6] As these advances approach maturity, a key outstanding question is how to use them to extract estimates as efficiently as possible when simulations are unsteady and chaotic and discretization methods are not significantly constrained by explicit stability limits.

The resolution of the turbulent—and therefore chaotic—nature of the underlying flows is a primary complication for estimation using these more complex simulations. For chaotic systems, estimation of long-time behavior is challenging, because the ordinary differential equations (ODEs) to which they can be reduced have limited predictability.[7] Of the general class of chaotic systems, a subset is ergodic systems, whose long-term states are drawn from a stationary distribution, independent of initial conditions (ICs).[8] For ergodic chaotic problems, we frequently want to quantify the unique infinite-time average of some instantaneous quantity of interest of the system

$$J_\infty = \lim_{T \to \infty} \frac{1}{T} \int_0^T g(\mathbf{u}(t)) \mathrm{d}t, \tag{1}$$

where g is the instantaneous output functional and the state $\mathbf{u}(t)$ is governed by a dynamical system of the form

$$\frac{\mathrm{d}\mathbf{u}}{\mathrm{d}t} = f(\mathbf{u}), \tag{2}$$

with a given initial condition (IC), $\mathbf{u}(0) = \mathbf{u}_{\mathrm{IC}}$.

Often, the complexity of chaotic systems of interest is high, and accordingly, the cost of an accurate computational estimate of J_∞ becomes formidable.[9–12] As the scale, cost, and complexity of a computational simulation increase, efficient discretization methods become critical for accurately estimating quantities of interest.

Understanding the error in approximations of J_∞ is nontrivial because statistical errors (errors due to a finite-time approximation)

and discretization error (error due to a numerical approximation of solutions) are always simultaneously present. In the largest direct numerical simulation (DNS) and large eddy simulation (LES) cases, for example, it is typical to fix the period of simulation time over which an average is performed at some large number of characteristic times and validate that discretization error converges as expected, assuming negligible sampling errors.[13–16] A recent work has sought to quantify the effect of statistical error more robustly, using a turbulent flow theory,[17] advanced spatiotemporal statistical post-processing methods,[18] statistical windowing techniques,[19] or by extending the concept of Richardson extrapolation to chaotic flows using auto-regressive models and Bayesian methods.[20] The latter work is notable for its use to estimate the statistical errors in the DNS of a high-$\Re$ turbulent channel flow.[21]

The objective of this paper is to investigate the behavior of statistical and discretization errors as a function of the computational cost for ergodic systems. Following a similar approach of Oliver et al.,[20] we propose a simple error model for finite-time, discrete approximations of infinite-time averages on attractors. Using the Lorenz system as an example, we demonstrate that the discretization error converges as the time step size decreases. However, if we denote by "sampling time" the amount of simulation time on which the integral is taken (after discarding spin-up), we show that error does not increase exponentially with sampling time as might be expected from classical numerical analysis but rather asymptotes to a constant, time step-dependent value with respect to the sampling time. Furthermore, for a given computational cost (e.g., the number of timsteps), an optimal choice of discretization (i.e., time step) exists that minimizes the expected error in simulation, when accounting for both the effects of discretization error and sampling error. We show that this optimal choice results in a convergence rate with respect to computational cost that is bounded by the sampling convergence rate with a minor impact from the discretization order of accuracy. Finally, we consider the implications of spin-up time (i.e., unsampled time needed to arrive at the stationary distribution) and parallelism on the optimal error. We develop a method for estimating transient-related errors and then evaluate optimal choices incorporating the results.

II. PROPOSED ERROR MODEL ON THE ATTRACTOR

To approximate J_∞, we compute finite-time, discrete estimates of the outputs of interest of the true system as

$$J_{T,hp} = \frac{1}{T_s} I_{t_0}^{t_0+T_s}(g_h p(\mathbf{u}_{hp}(t))), \tag{3}$$

where the notation $I_a^b(\cdot)$ here represents the quadrature approximation of the integral $\int_a^b (\cdot)\mathrm{d}t$ of a quantity $(\cdot)$ between a and b. Here, we have made a discrete approximation of the state using an order-p discretization with a temporal grid with characteristic size $h = \Delta t$, where an order-p discretization is one for which the discretization error behaves as

$$\max_{t\in[0,t_0+T_s]} |g_h p(\mathbf{u}_{hp}(t)) - g(\mathbf{u}(t))| = \mathcal{O}(h^p), \tag{4}$$

when the discretization is applied to a well-posed (non-chaotic) system. Then we sample a discrete state over a finite sampling period, T_s, starting at some initial time t_0. We can define the error that is incurred as

$$e_{T,hp} = J_{T,hp} - J_\infty. \tag{5}$$

By introducing a third value,

$$J_T = \frac{1}{T_s} \int_{t_0}^{t_0+T_s} g(\mathbf{u}(t))\mathrm{d}t, \tag{6}$$

we can re-write the error using an identity

$$e_{T,hp} = (J_{T,hp} - J_T) + (J_T - J_\infty) = e_{hp} + e_T. \tag{7}$$

Here, we define the "discretization error" and "sampling error," respectively, as

$$e_{hp} \equiv J_{T,hp} - J_T, \tag{8}$$

$$e_T \equiv J_T - J_\infty. \tag{9}$$

We can take an absolute value of both sides of (7), followed by a manipulation using the triangle inequality

$$|e_{T,hp}| = |e_{hp} + e_T|$$
$$\leq |e_{hp}| + |e_T|. \tag{10}$$

Thus, the total error incurred by the approximation is bounded by the sum of the absolute discretization and sampling errors. Next, we define the attractor of the operator f, $\mathcal{A}$, as the set of long-term states toward which all trajectories converge independently of the initial condition.[22] We can define the expectation $\mathbb{E}_\mathcal{A}[\phi(\mathbf{u}_0)]$ for a generic function ϕ as the expectation taken over all the trajectories that can result from starting from points on the attractor, $\mathcal{A}$,

$$\mathbb{E}_\mathcal{A}[\phi] = \frac{1}{|\mathcal{A}|} \int_{\mathbf{u}_0\in\mathcal{A}} \phi(\mathbf{u}_0)\mathrm{d}\mathbf{u}_0. \tag{11}$$

For the case in question, we will be considering either

$$\phi(\mathbf{u}_0) = \frac{1}{T_s}\left| \int_{t_0}^{t_0+T_s} g(\mathbf{u}(t))\mathrm{d}t - J_\infty \right|$$

or

$$\phi(\mathbf{u}_0) = \frac{1}{T_s}\left| I_{t_0}^{t_0+T_s}(g_h p(\mathbf{u}_{hp}(t))) - \int_{t_0}^{t_0+T_s} g(\mathbf{u}(t))\mathrm{d}t \right|,$$

with, for these examples, $\mathbf{u}(t_0) = \mathbf{u}_0 \in \mathcal{A}$. Given these definitions, we can now take the expectation of (10), giving

$$\mathbb{E}_\mathcal{A}[|e_{T,hp}|] \leq \mathbb{E}_\mathcal{A}[|e_{hp}|] + \mathbb{E}_\mathcal{A}[|e_T|], \tag{12}$$

by linearity.

From here, we propose asymptotic forms for the two right-hand side terms in (12). Consider the definition of e_T in (9)

$$e_T = \frac{1}{T_s} \int_{t_0}^{t_0+T_s} g(\mathbf{u}(t))\mathrm{d}t - J_\infty. \tag{13}$$

Assuming that we choose t_0 such that each $\mathbf{u}_0$ is effectively an independent sample from the attractor's stationary distribution, then the quantity $g(\mathbf{u}(t))$ is a random variable drawn from a stationary distribution. The states of ergodic systems, in general, are not independent in time, but as long as the system has satisfactorily strong mixing properties, the central limit theorem (CLT) can be applied to finite time

averages of its outputs. This is the case whenever the condition of α-mixing is met,[23,24] which has been shown for the Lorenz system.[25] Thus, we can write e_T as

$$e_T \sim \mathcal{N}\left(0, \left(\sqrt{\frac{\pi}{2}} A_0 T_s^{-1/2}\right)^2\right), \tag{14}$$

with a constant coefficient A_0, where $\mathcal{N}(\mu, \sigma^2)$ gives the normal distribution with mean μ and variance σ^2. If we take the absolute value of this random variable, the result is a halfnormal distribution

$$|e_T| \sim \mathcal{H}\left(\left(\sqrt{\frac{\pi}{2}} A_0 T_s^{-1/2}\right)^2\right), \tag{15}$$

where $\mathcal{H}(\sigma^2)$ gives a halfnormal distribution such that $|X| \sim \mathcal{H}(\sigma^2)$ when $X \sim \mathcal{N}(0, \sigma^2)$. The expectation of the half-normal distribution is well defined, allowing

$$\mathbb{E}_{\mathcal{A}}[|e_T|] \approx A_0 T_s^{-1/2}, \tag{16}$$

as T_s goes to infinity.

Now consider the use of a time-stepping method to give a discrete approximation $\mathbf{u}_{hp}(t_n)$ of $\mathbf{u}(t_n)$ for each $t_n = n(\Delta t)$. Following the classical analysis,[26] we might expect that the discretization error should take a form

$$|e_{hp}| \approx C_p\left(\frac{\exp(\Lambda T_s) - 1}{\Lambda}\right)(\Delta t)^p. \tag{17}$$

This analysis is based on bounding the growth of local truncation error at each time step by the Lipschitz constant, Λ, of the underlying system, with C_p being a constant parameter that depends on the choice of method. However, Viswanath showed[27] that the global error could be modeled by a form

$$|e_{hp}| \approx E(T_s; p)(\Delta t)^p, \tag{18}$$

where $E(T_s; p)$ could be bounded by a constant for some nonlinear but non-chaotic systems that are exponentially stable. While this result has not been extended to an ergodic system, the expected convergence onto the attracting set suggests a bound of the form

$$\mathbb{E}_{\mathcal{A}}[|e_{hp}|] \approx C_p(\Delta t)^p. \tag{19}$$

As our results in Sec. III will show, (19) is a good description of the expected discretization error.

Thus, taking (12), (16), and (19), we assume a bound of the form

$$\mathbb{E}_{\mathcal{A}}[|e_{T,hp}|] \leq e_{\text{model}} = C_q(\Delta t)^q + A_0 T_s^{-r}$$
$$= C_q(\Delta t)^q + A_0 N_s^{-r}(\Delta t)^{-r}, \tag{20}$$

that bounds $\mathbb{E}_{\mathcal{A}}[|e_{T,hp}|]$ when Δt is small enough and T_s is large enough to satisfy the asymptotic assumptions. Here, we have used $T_s = N_s \Delta t$, where N_s is the number of timesteps computed for sampling; additionally, q refers to the observed discretization convergence behavior, which, in practice, may differ from p due to numerical cancelations or if the solutions of the system are insufficiently regular. Similarly, r is an observed sampling convergence rate coefficient, which we expect to be 1/2 asymptotically under the CLT.

III. EVALUATION OF THE PROPOSED ERROR MODEL ON THE LORENZ SYSTEM

In this section (Sec. III), we will fit numerical results for the Lorenz system to determine q, r, C_q, and A_0 and show that this model is representative of the observed behavior. The Lorenz system is given by:[28]

$$\frac{d\mathbf{u}}{dt} = f(\mathbf{u}; \boldsymbol{\alpha}) = \begin{pmatrix} \alpha_0(u_1 - u_0) \\ u_0(\alpha_1 - u_2) - u_1 \\ u_0 u_1 - \alpha_2 u_2 \end{pmatrix}, \tag{21}$$

where $\mathbf{u} = [u_0, u_1, u_2]^\top$ and $\boldsymbol{\alpha} = [\alpha_0, \alpha_1, \alpha_2]^\top$. The Lorenz system is known to be chaotic for the classic Lorenz parametrization:[29] $\boldsymbol{\alpha} = [10, 28, 8/3]$, which is used everywhere in this text. For the output, we choose $g(\mathbf{u}) = u_2$. We consider a set of explicit methods: forward Euler (FE, $p = 1$), third-order Runge–Kutta (RK3, $p = 3$), and fourth-order Runge–Kutta (RK4, $p = 4$). In all of these methods, we expect asymptotic convergence of $J_{T,hp}$ to J_T to be at least $\mathcal{O}(\Delta t^p)$ for non-chaotic systems.[30] The stability limit for each of these methods is $\Delta t \gtrsim 10^{-1}$, which is above the limits of asymptotic convergence for all of the methods here; tests with implicit methods indicated similar error trends at higher cost and are, therefore, omitted.

For any given discrete instance, we will start the simulation at an initial state at $t = 0$ that is sampled randomly from a normal distribution

$$\mathbf{u}_{\text{init}} \sim \begin{pmatrix} \mathcal{N}(1.0, 5.0^2) \\ \mathcal{N}(1.0, 5.0^2) \\ \mathcal{N}(1.0, 5.0^2) \end{pmatrix}. \tag{22}$$

To guarantee that the initial sampling state $\mathbf{u}_0$ at t_0 is on the attractor (as well as further guaranteeing the independence from the other Monte Carlo instances), we evolve the state of any given Lorenz system discretization from its starting state $\mathbf{u}_{\text{init}}$ for $t_0 = 100$ before proceeding to the sample; we refer to the process of evolving the solution until it is on the attractor as "spin-up." Then, we evolve the state over the next T_s, during which we integrate and compute (3) using the same numerical integration scheme that was used for the state itself.

To approximate $e_{T,hp}$, we must first estimate J_∞ by a reference value J_{ref}. J_{ref} is calculated using an ensemble mean of $J_{T,hp}$ over $M_{\text{ens}} = 512^2$ instances of the Lorenz system. Each instance is started from a different $\mathbf{u}_{\text{init}}$ as given in (22) and simulated using RK4 with $\Delta t = 17.7 \times 10^{-6}$ and $T_s = 6646.9$. The resulting J_{ref} is

$$J_{\text{ref}} = 23.549\,916 \pm 0.000\,074, \tag{23}$$

with a 95% confidence estimate based on the ensemble mean estimator.

The computation of J_{ref} allows us to estimate errors $e_{T,hp} \approx J_{T,hp} - J_{\text{ref}}$. For a given Δt, T_s pair, we then approximate $\mathbb{E}_{\mathcal{A}}[|e_{T,hp}|]$ using a Monte Carlo method over $M = 10\,000$ independent instances of the discrete system, each started from initial states drawn from (22) and spun-up to independent sampling starting points on the attractor $\mathbf{u}_0^{(m)}$,

$$\mathbb{E}[|e_{T,hp}|] \approx \frac{1}{M} \sum_{m=1}^{M} \left| J_{T,hp}\left(\mathbf{u}_0^{(m)}\right) - J_{\text{ref}} \right|. \tag{24}$$

FIG. 1. Expected relative error as a function of Δt for forward Euler discretization of the Lorenz equations. Nonlinear least squares fit based on $N_s = 10^6$ data.
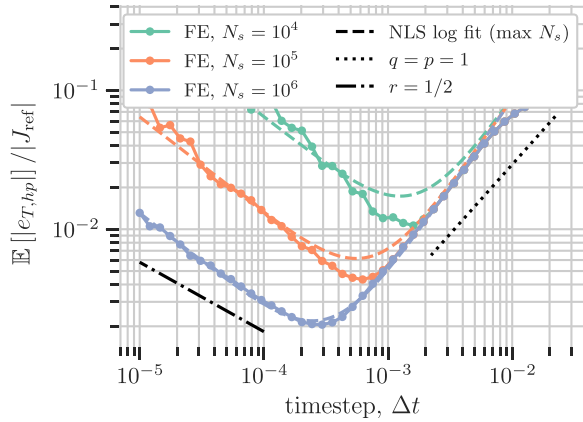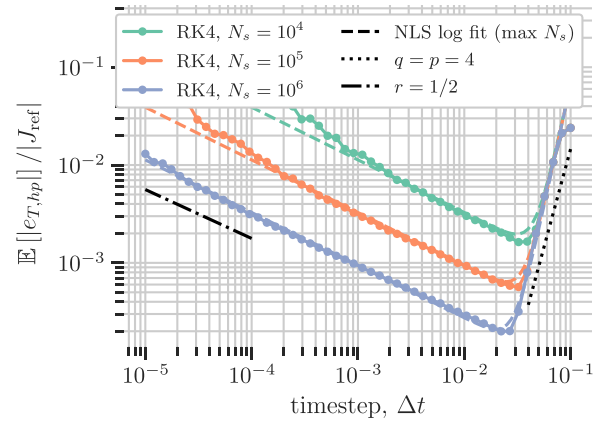

FIG. 3. Expected relative error as a function of Δt for RK4 discretization of the Lorenz equations. Nonlinear least squares fit based on $N_s = 10^6$ data.

In Figs. 1–3, we compare the results of simulations with the FE, RK3, and RK4 discretizations with different values of N_s. In these figures, T_s scales with Δt for a given N_s, so the T_s values on the x-axis will vary between lines on the plot. The fits shown are computed with truncated data in order to attempt to eliminate non-convergent data at small T_s or large Δt; the limits used for truncation are found in Table I. The results of the nonlinear least squares fits for $N_s = 10^4$, 10^5, and 10^6 are given in Table II. In the table, we observe that $r \to 1/2$ as the discretization error is reduced, either by increasing N_s or by pushing p higher. These figures demonstrate that (19) has explanatory value, as the errors in the discretization-dominated region collapse independently of T_s. It is also worth noting that Table II demonstrates higher-than-expected discretization error convergence rates for FE and RK4. In Fig. 4, we can examine the sampling error behavior between discretization methods for a single shared choice of N_s. Here, we can see that the sampling error effects on the left-hand side of the plot collapse independently of the discretization method. This indicates that the statistical effects are properties of the dynamical system, not artifacts of the discretization, as we might expect in the limit as

TABLE I. Fit boundaries for nonlinear least squares fits.

Method	Δt_{max}	$T_{s,min}$
FE	5.0×10^{-3}	1.0
RK3	5.0×10^{-2}	1.0
RK4	9.0×10^{-2}	1.0

$\Delta t \to 0$. We note that Table II shows some N_s-dependence in the fit values of r; we expect that this is an artifact of the fit process, in which discretization effects and statistical effects are intermingled. As N_s or q become small, the region in which the convergent CLT behavior is the dominant effect becomes smaller, and for these cases, the value of r detected is subject to error and deviation from the expected rate $r = 1/2$.

TABLE II. Values of error model coefficients computed from nonlinear least squares fits to Monte Carlo study data.

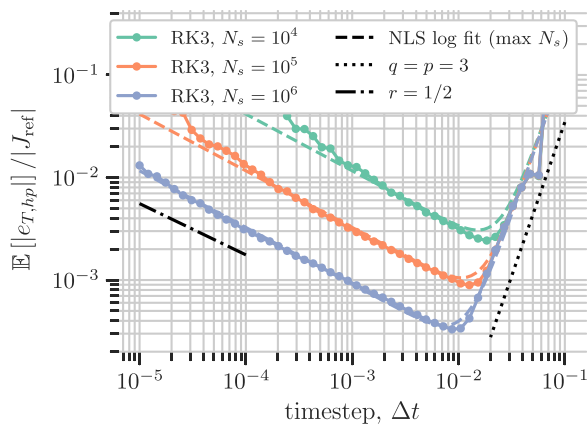|	FE	RK3	RK4
	(a) $N_s = 10^4$		
A_0	2.19	1.74	1.63
r	0.975	0.721	0.683
C_q	4995	942	85 900
q	1.65	2.70	4.83
	(b) $N_s = 10^5$		
A_0	1.94	1.50	1.41
r	0.820	0.648	0.620
C_q	1410	1310	96 100
q	1.40	2.76	4.84
	(c) $N_s = 10^6$		
A_0	1.52	0.978	0.918
r	0.693	0.553	0.538
C_q	714.6	2740	165 000
q	1.273	2.96	5.02

FIG. 2. Expected relative error as a function of Δt for RK3 discretization of the Lorenz equations. Nonlinear least squares fit based on $N_s = 10^6$ data.
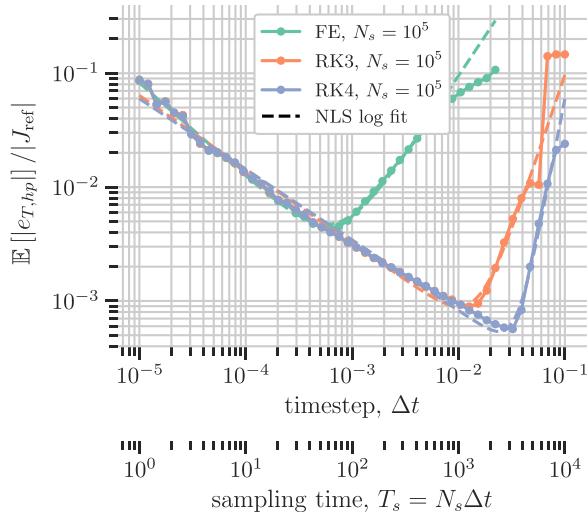
FIG. 4. Expected relative error as a function of Δt for discretizations of the Lorenz equations.

Finally, we attempt to compare the computational costs across the various discretizations. In this case, the number of timesteps N_s is not a good proxy for fixed cost, since the computation time for a time step will vary between methods. Instead, we now fix U_s, the total number of evaluations of the right-hand side f used in sampling timesteps. For the explicit schemes used in this work, we will have p right-hand side evaluations (e.g., forward Euler has $p = 1$ right-hand side evaluations) and, thus, $U_s = pN_s$. In Fig. 5, we can see the effect of changing Δt at fixed sampling cost U_s across discretizations. The error that can be achieved with the Runge–Kutta methods is lower than that of the forward Euler scheme, a factor of 4.8 improvement in the error from FE to RK4. However, the best-case improvement for going from third-order to fourth-order Runge–Kutta schemes is the only factor of about 1.4. Moreover, the results show that to achieve the lowest possible
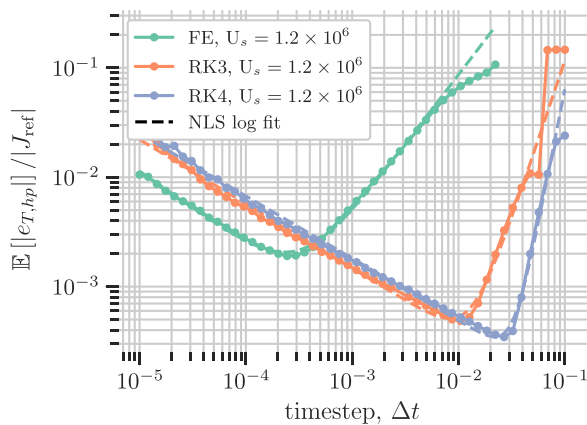

FIG. 5. Expected percent error as a function of Δt for discretizations of the Lorenz equations at a number of sampling residual evaluations. All fits evaluated at $U_s = 1.2 \times 10^6$.

error, the optimal time step will be discretization dependent. We investigate this further in Sec. IV.

IV. OPTIMAL TIMESTEPPING ON THE ATTRACTOR

We now study the implications of the error model (20), specifically seeking to understand the convergence of the error with respect to computational effort. In this analysis, we will assume that $r = 1/2$.

Consider a non-dimensional form of the error model in which the error is normalized by the standard deviation of the instantaneous output σ_g, and the timescales Δt and T_s are normalized by decorrelation time T_d. The decorrelation time relates the amount of variance from independent draws from the distribution on the attractor and the amount of variance in the finite-time mean estimators based on the correlated output signal, given by the relation:[31]

$$\text{Var}[J_T] = \frac{T_d}{T_s}\sigma_g^2. \tag{25}$$

Furthermore, combining (15) and (25) allows us to write

$$A_0 = \sqrt{\frac{2}{\pi}}\sigma_g T_d^{1/2}. \tag{26}$$

In general, T_d is hard to estimate accurately; this is a crux of the work of Oliver *et al.*[20] In our formulation of the error model, we identify A_0, which avoids outright estimation of T_d. However, for the purpose of understanding the behavior of the error, T_d is an intrinsic timescale, which can be used to normalize Δt and T_s.

The resulting non-dimensional form of the error model is

$$\frac{e_{\text{model}}}{\sigma_g} = \frac{C_q T_d^q}{\sigma_g}\left(\frac{\Delta t}{T_d}\right)^q + \sqrt{\frac{2}{\pi}}\left(\frac{T_s}{T_d}\right)^{-\frac{1}{2}}. \tag{27}$$

We can also write the optimizers and optimal value of (27) in terms of the non-dimensional variables. These are given by

$$\left(\frac{\Delta t}{T_d}\right)_{\text{opt}} = \left(\frac{1}{2\pi}\right)^{\frac{1}{2q+1}}\left(\frac{qC_q T_d^q}{\sigma_g}\right)^{-\frac{2}{2q+1}}N_s^{-\frac{1}{2q+1}},$$

$$\left(\frac{T_s}{T_d}\right)_{\text{opt}} = \left(\frac{1}{2\pi}\right)^{\frac{1}{2q+1}}\left(\frac{qC_q T_d^q}{\sigma_g}\right)^{-\frac{2}{2q+1}}N_s^{\frac{2q}{2q+1}}, \tag{28}$$

$$\left(\frac{e_{\text{model}}}{\sigma_g}\right)_{\text{opt}} = \left(\frac{1}{2\pi}\right)^{\frac{q}{2q+1}}\left(2+\frac{1}{q}\right)\left(\frac{qC_q T_d^q}{\sigma_g}\right)^{\frac{1}{2q+1}}N_s^{-\frac{q}{2q+1}}.$$

In terms of convergence with respect to sampling costs, the error model will scale at best as

$$\left(\frac{e_{\text{model}}}{\sigma_g}\right)_{\text{opt}} \sim N_s^{-\frac{q}{2q+1}}.$$

In the limit as $q \to \infty$, the rate $q/(2q+1) \to 1/2$: the CLT limits the convergence rate. Table III gives the rates of convergence (28) for various values of q.

Using the reference simulation, we can also find

$$\begin{aligned}\text{Var}[J_T] &\approx \text{Var}[J_{T,hp}] = 1.1692 \times 10^{-4}, \\ \sigma_g^2 &\approx \hat{\sigma}_g^2 = 74.348\,04 \pm 0.000\,18,\end{aligned} \tag{29}$$

TABLE III. Convergence rates for combined error with respect to sampling timesteps implied in (28) at common high-order discretization error convergence rates.

q	1	2	3	4	5	$\cdots$	∞
$\frac{q}{2q+1}$	1/3	2/5	3/7	4/9	5/11	$\cdots$	1/2

where $\hat{\sigma}_g$ is an estimate of the standard deviation of g. Together, these allow us to estimate

$$T_d \approx 1.0170 \times 10^{-2},$$
$$\sigma_g \approx 8.6225. \tag{30}$$

With these values, we can plot the non-dimensional error model with fixed $r = 1/2$, which is given for $N_s = 10^5$ in Fig. 6.

　　We now consider the implications of these results for increasing N_s. To focus solely on control of the discretization error, increases in N_s can be used to refine $\Delta t = T_s/N_s$, with T_s fixed. On the other hand, to focus solely on controlling the sampling error, $T_s = N_s\Delta t$ can be increased, holding Δt fixed. In Fig. 7, the two approaches are compared with the optimal use of resources. In orange is the discretization error control strategy. In this approach, the simulations converge at a high-order rate in N_s toward the optimal error behavior; once the error reaches this optimum, however, it asymptotes to a constant: statistical errors limit the estimation of $J_{T,hp}$. On the other hand, the sampling error control approach is shown in blue. In this approach, the central limit convergence rate of 1/2 is initially achieved until the error asymptotes to a constant: discretization errors limit the estimation of $J_{T,hp}$. In other words, this latter case represents having increasingly many significant figures of certainty about an answer that is significantly offset from the true quantity of interest, J_∞. In the literature for large simulations, discussed in the introduction, simulations tend to be planned using either the discretization or statistical error control approach. What (20) implies and Fig. 7 demonstrates is that, in fact, there is a
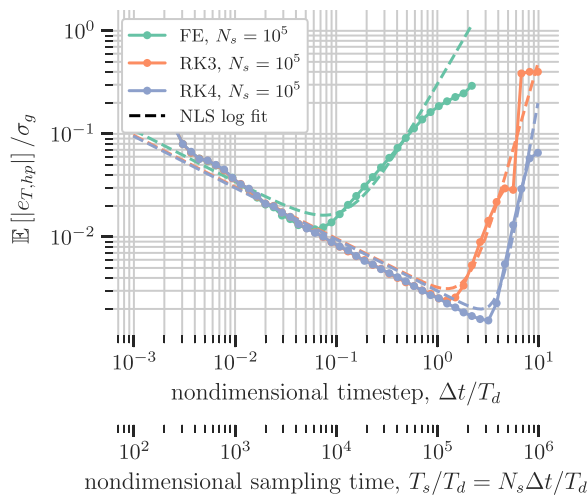

FIG. 6. Expected non-dimensional error as a function of the non-dimensional time step for discretizations of the Lorenz equations. $r = 1/2$ assumed.
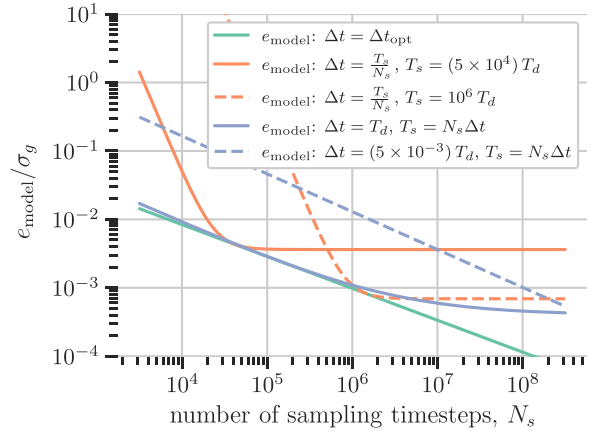

FIG. 7. Refinement study comparison for fixed Δt, fixed T_s, and optimized Δt and T_s using RK3 discretization to compute the expectation of the Lorenz system output $g = u_2$.

particular optimal scheme in which Δt and T_s are simultaneously varied that will extract the most accurate estimate of J_∞ as N_s increases.

V. INVESTIGATION OF THE GLOBAL DISCRETIZATION ERROR MODEL

　　In this section, we show that our simulations of chaotic, ergodic ODEs are consistent with a bounded relationship between the local and global discretization errors. Consider an estimate of the global error based on N_s timesteps

$$e_{hp} \approx \frac{1}{N_s}\sum_{n=0}^{N_s}\sum_{\eta=n}^{N_s}\mathcal{G}(t_\eta, t_n) \circ \mathbf{e}_{\mathrm{LT},p}^{(n)}, \tag{31}$$

where

$$\mathbf{e}_{\mathrm{LT},p}^{(n)} \equiv \mathbf{u}_{hp}(t_{n+1}) - \mathbf{u}_\star(t_{n+1}), \tag{32}$$

and $\mathbf{u}_\star(t_{n+1})$ is the exact solution integrated from $\mathbf{u}_{hp}(t_n)$ through Δt

$$\mathbf{u}_\star(t_{n+1}) = \mathbf{u}_{hp}(t_n) + \int_{t_n}^{t_{n+1}} f(\mathbf{u}_\star(t))\mathrm{d}t. \tag{33}$$

In (31), we have assumed that the error from any given local state perturbation is propagated forward in time by the dynamics, before being transformed into an error in the output; this process is captured by an operator $\mathcal{G}$. Because the effect of local error propagates forward and not backward in time, $\mathcal{G}(t, t_n) = 0$ for $t < t_n$, and moreover, we assume that due to ergodicity $\mathcal{G}(t, t_n) = 0$ when $t - t_n \gtrsim T_d$, where T_d is the decorrelation time associated with the attractor. This allows us to write

$$e_{hp} \approx \frac{1}{N_s}\sum_{n=0}^{N_s}\sum_{\eta=n}^{n+T_d/\Delta t}\mathcal{G}(t_\eta, t_n) \circ \mathbf{e}_{\mathrm{LT},p}^{(n)}. \tag{34}$$

　　Now, we assume that a constant $\mathcal{G}_{\max}$ exists such that

$$|\mathcal{G}(t_\eta, t_n) \circ \mathbf{v}| \leq \mathcal{G}_{\max}||\mathbf{v}||_\infty, \tag{35}$$

for all $t_n, t_\eta \in \mathbb{R}$, and $\mathbf{v} \in B(\mathbf{u}(t_n)) \subset \mathbb{R}^d$, where $B(\mathbf{u})$ is the set of states possible by perturbation of $\mathbf{u}$ that remain in the basin of attraction of the attractor $\mathcal{A}$ of f. When this is the case, we can create a bound on the magnitude of e_{hp}

$$|e_{hp}| \leq \frac{T_d}{\Delta t} \mathcal{G}_{max} \frac{1}{N_s} \sum_{n=0}^{N_s} ||\mathbf{e}_{LT,p}^{(n)}||_\infty$$

$$\leq \frac{T_d}{\Delta t} \mathcal{G}_{max} \max_n ||\mathbf{e}_{LT,p}^{(n)}||_\infty. \tag{36}$$

We now attempt to bound the value of $\mathcal{G}_{max}$ for the Lorenz system by approximating the local truncation error. To make an estimate, we compute both the solution at the next time step as well as a surrogate for the true solution at each time step: $\mathbf{u}_{hp}(t_{n+1})$ and $\tilde{\mathbf{u}}_\star(t_{n+1})$, where the former is computed with one time step of the method of interest and the latter is always computed with the highest available accuracy method, RK4, and subdividing $t \in [t_n, t_{n+1}]$ into ten consecutive timesteps rather than one. Both $\mathbf{u}_{hp}(t_{n+1})$ and $\tilde{\mathbf{u}}_\star(t_{n+1})$ are always advanced from $\mathbf{u}_{hp}(t_n)$. This allows us to estimate $\mathbf{e}_{LT,p}^{(n+1)}$ locally

$$\mathbf{e}_{LT,p}^{(n+1)} \approx \tilde{\mathbf{e}}_{LT,p}^{(n+1)} = \mathbf{u}_{hp}(t_{n+1}) - \tilde{\mathbf{u}}_\star(t_{n+1}). \tag{37}$$

In Fig. 8, we characterize the convergence of local error estimates. Computations are run with $T_s = 100$ and $t_0 = 100$ fixed, varying Δt. At each time step, the local truncation error is estimated by computing (37). The figure shows the computed $\max_n ||\tilde{\mathbf{e}}_{LT,p}^{(n)}||_\infty$ and demonstrates that the expected rate of $(p+1)$ is nearly exactly achieved.

Using (36), we can estimate a bounding value for $\mathcal{G}_{max}$ by

$$\mathcal{G}_{max} \geq \frac{\mathbb{E}[|e_{hp}|]}{\max_n ||\mathbf{e}_{LT,p}^{(n)}||_\infty} \frac{\Delta t}{T_d} = \frac{C_q \Delta t^q}{c_p \Delta t^{p+1}} \frac{\Delta t}{T_d}, \tag{38}$$

where c_p is the leading truncation error coefficient fit in Fig. 8 and C_q and q are taken from Table II. Of course when $q > q_{theory} = p$, there will be Δt dependence. However, as (38) requires that the discretization error has an asymptotic behavior, we will only consider Δt in the asymptotic convergence regions given in Table I to compute $\mathcal{G}_{max}$. In

FIG. 9. Estimation of bounding value $\mathcal{G}_{max}$.

Fig. 9, we show the values of the right-hand side quantity in (38), which allow us to make an estimate

$$\mathcal{G}_{max} \approx 4.3. \tag{39}$$

Next, we use classical truncation error estimates[26] to relate the discretization error to properties of the solution. We will assume that the local truncation error is bounded by a form

$$\max_n ||\mathbf{e}_{LT,p}^{(n)}||_\infty \leq \frac{C_{LT}}{(p+1)!} \left\| \frac{d^{p+1}\mathbf{u}}{dt^{p+1}} \right\|_\infty \Delta t^{p+1}, \tag{40}$$

where C_{LT} is a local truncation constant term dependent on the numerical method and the $|| \cdot ||_\infty$ in this context refers to the maximum value in time of the inf-norm of a vector-valued, time-dependent quantity $(\cdot)$. The derivatives of $\mathbf{u}(t)$ can be computed by evaluating $f(\mathbf{u})$ and its derivatives using solutions from a reference RK4 solution of the Lorenz system with $T_s = 1000$, $t_0 = 100$, and $\Delta t = 10^{-4}$. Norms of the derivatives are shown in Fig. 10. The resulting values of C_{LT} that can now be derived by fitting the asymptotic behavior in Fig. 8 can be found in Table IV. The result of these estimates is that we
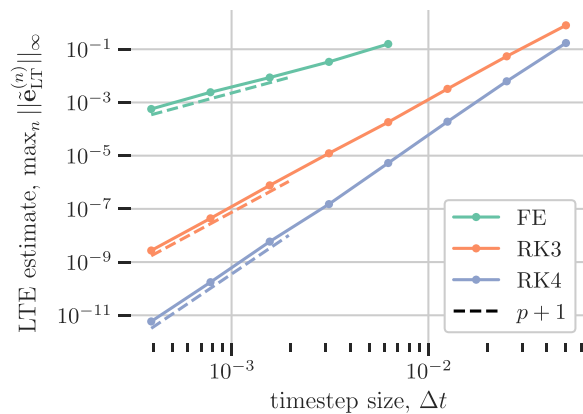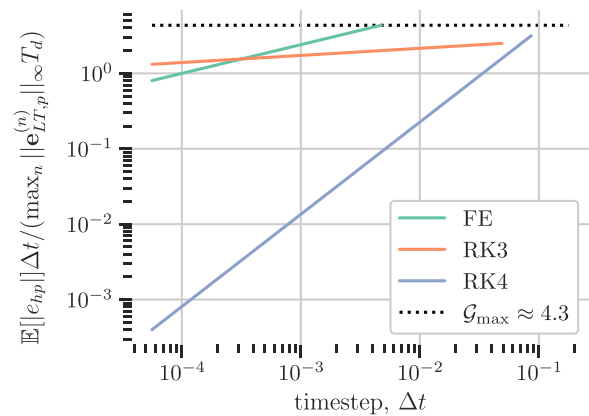

FIG. 8. Convergence of estimated local truncation error with respect to Δt. Fits to $c_p \Delta t^{p+1}$ shown (with offset for presentation).
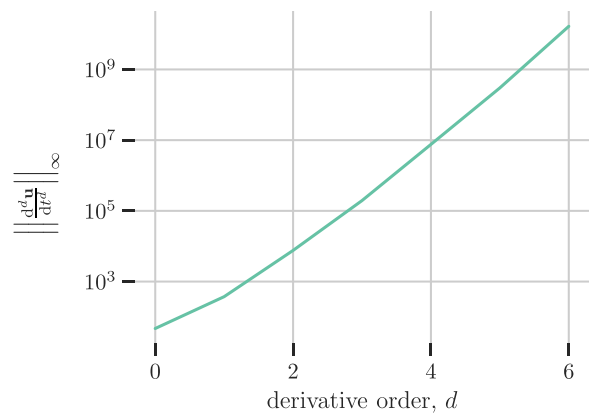

FIG. 10. Norm of analytic derivatives of $\mathbf{u}$ computed on the attractor of f. State $\mathbf{u}$ computed with RK4 at $\Delta t = 10^{-4}$ and $T_s = 1000$ after discarding $t_0 = 100$.

TABLE IV. Rate and coefficient fit for convergence of local truncation error of the discrete Lorenz system. $C_{LT}\|\frac{d^{p+1}\mathbf{u}}{dt^{p+1}}\|_\infty$ estimated by $c_p(p+1)!$ using c_p fit from Fig. 8.

p	rate (observed)	$C_{LT}\|\frac{d^{p+1}\mathbf{u}}{dt^{p+1}}\|_\infty$	C_{LT}
1	2.00	7.61×10^3	7.33
3	4.02	3.28×10^6	156
4	4.93	4.50×10^7	76.5

can reliably bound the global error of a dynamical system as an accumulation of the local errors over a region of correlation.

We now want to consider how the global error behavior demonstrated here might extrapolate to more complicated systems by evaluating the spectral behavior of the Lorenz system. Using a discrete Fourier transform (DFT) with a Hann window function,[32] we perform a spectral analysis on the states of the Lorenz system with sampling time $T_s = 1000$, $t_0 = 100$, and $\Delta t = 10^{-3}$. The resulting spectrum can be found in Fig. 11. The Lorenz system tends to have the most content in the frequencies with $f \lesssim 10^1$ with a region of exponential decay in the range $1 \lesssim f \lesssim 300$. On scales with $f \gtrsim 300$, machine precision plateaus are observed and omitted here.

The fact that the Lorenz spectrum is an exponentially decreasing function of frequency f makes the use of high-order methods theoretically appealing for the spectral convergence of hp-refinement strategies.[33] Unfortunately, the effect of statistical error in (28) limits the impact of this exponential decay, such that the benefits of higher-order discretization methods are limited compared to their steady-state and non-chaotic application. The convergence to the central limit rates can be seen in Fig. 12, which shows the convergence of (28) with the total sampling cost. The effect of increasing order improves the convergence rate in (28) toward the CLT-implied asymptotic rate of $-1/2$, as well as decreasing the value of the leading constant and the error never achieves the spectral rates possible with hp-refinement in the steady case. Nevertheless, the cost to achieve a given amount of error in expectation—in terms of function evaluations—is significantly less
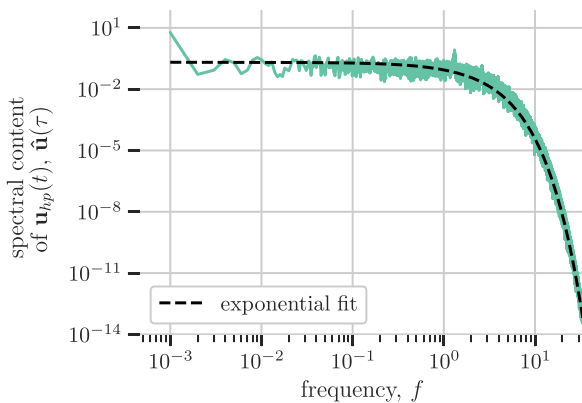

FIG. 12. Convergence of optimal error with sampling costs for FE, RK3, and RK4 discretizations of the Lorenz output $g = u_2$. Asymptotic $-1/2$ rate implied by the central limit theorem is shown.

with higher-order methods. Managing to achieve 1% non-dimensional error in expectation is possible with RK4 at a cost ten times less than would be possible using FE; this factor grows larger than 100 when the tolerance is tightened to 10^{-4}.

VI. IMPACT OF ENSEMBLE AVERAGING AND SPIN-UP

In this section, we will consider how the error behaves when ensemble averaging (over multiple parallel instances) and when spin-up effects are present.

A. Ensemble averaging on the attractor

Sampling error can be reduced at a fixed wall clock time by ensemble averaging across multiple parallel processes.[34] Consider a Monte Carlo approach to approximate J_∞ with a set of M_{ens} independent realizations

$$J_{MC} = \frac{1}{M_{ens}} \sum_{m=1}^{M_{ens}} J_{T,hp}^{(m)}. \quad (41)$$

We can write a modified version of (20) to approximate the error that we expect in the Monte Carlo estimator in (41)

$$\mathbb{E}[|J_{MC} - J_\infty|] \approx e_{model,MC} = C_q(\Delta t)_{MC}^q + \frac{A_0}{\sqrt{M_{ens}}} T_{s,MC}^{-r}, \quad (42)$$

with an equivalent non-dimensional version, assuming $r \to 1/2$

$$\left(\frac{e_{model}}{\sigma_g}\right)_{MC} = \frac{C_q T_d^q}{\sigma_g}\left(\frac{\Delta t}{T_d}\right)^q + \sqrt{\frac{2}{\pi}} M_{ens}^{-\frac{1}{2}}\left(\frac{T_s}{T_d}\right)^{-\frac{1}{2}}, \quad (43)$$

and an optimum given by

$$\left(\frac{e_{model}}{\sigma_g}\right)_{MC,opt} = \left(\frac{1}{2\pi}\right)^{\frac{q}{2q+1}}\left(2 + \frac{1}{q}\right)\left(\frac{qC_q T_d^q}{\sigma_g}\right)^{\frac{1}{2q+1}} M_{ens}^{-\frac{q}{2q+1}} N_s^{-\frac{q}{2q+1}}, \quad (44)$$

at

FIG. 11. Fourier spectrum of $\mathbf{u}(t)$. Computed with DFT using the Hann window function on data from RK4 discretization of the Lorenz system with $T_s = 1000$, $t_0 = 100$, and $\Delta t = 10^{-3}$. Gray dashed line: fit assuming $|\hat{\mathbf{u}}(f)| \approx \exp(-af + b)$ with $a = 0.872$ and $b = 2.58$.
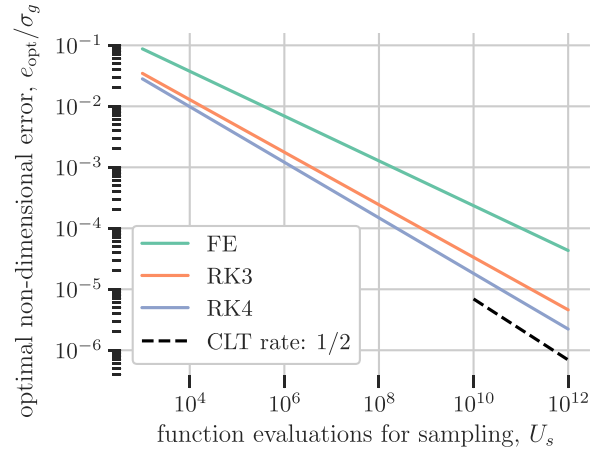
$$\left(\frac{\Delta t}{T_d}\right)_{\text{opt}} = \left(\frac{1}{2\pi}\right)^{\frac{1}{2q+1}} \left(\frac{qC_q T_d^q}{\sigma_g}\right)^{-\frac{2}{2q+1}} M_{\text{ens}}^{-\frac{1}{2q+1}} N_s^{-\frac{1}{2q+1}} \quad (45)$$

and

$$\left(\frac{T_s}{T_d}\right)_{\text{opt}} = \left(\frac{1}{2\pi}\right)^{\frac{1}{2q+1}} \left(\frac{qC_q T_d^q}{\sigma_g}\right)^{-\frac{2}{2q+1}} M_{\text{ens}}^{-\frac{1}{2q+1}} N_s^{\frac{2q}{2q+1}}. \quad (46)$$

Equation (44) shows that, for finite values of q, the Monte Carlo method will have a mitigated return compared to its purely stochastic application as in Ref. 34; the optimal error scales as $M_{\text{ens}}^{-q/(2q+1)}$ as opposed to $M_{\text{ens}}^{-1/2}$. However, parallelization can achieve perfect scaling in the expected error, in the sense that the effect of running M_{ens} ensembles with N_s sampling timesteps each will have an equivalent error in expectation to simulate $M_{\text{ens}} N_s$ timesteps in series. As M_{ens} is varied on the set of optimal solutions, (45) and (46) indicate that the time step and sampling time should be adjusted with the same factor $M_{\text{ens}}^{-1/(2q+1)}$ to achieve perfect scaling.

B. Spin-up transient modeling

So far, we have considered the error and cost on the attractor, neglecting the impact of spin-up from $t = 0$ to $t = t_0$. This spin-up is necessary because simulations of ergodic systems invariably need some time for the state to proceed onto the attractor from the initial condition.

Consider $\mathbf{u}(t)$, a solution of the ergodic chaotic system f from an arbitrary initial condition $\mathbf{u}(0) = \mathbf{u}_{\text{IC}}$ in the basin of attraction of an attractor, $\mathcal{A}$. The existence of the attractor implies the nonlinear stability of the system, such that all $\mathbf{u}_{\text{IC}}$ will converge to trajectories on the attractor $\mathcal{A}$. Denote by $\mathbf{u}^{\mathcal{A}}(t)$ a trajectory that is on the attractor for all t and to which $\mathbf{u}(t)$ collapses as $t \to \infty$. The perturbation $\delta \mathbf{u}^{\mathcal{A}}(t) \equiv \mathbf{u}(t) - \mathbf{u}^{\mathcal{A}}(t)$ that describes the IC, therefore, exists in a stable subspace of perturbations to $\mathbf{u}^{\mathcal{A}}$ and can be associated with the negative Lyapunov exponents of the system. Thus, we can assume that such perturbations are governed asymptotically by

$$||\delta \mathbf{u}^{\mathcal{A}}(t)|| \lesssim \exp\left(-\frac{t}{T_\lambda}\right), \quad (47)$$

with T_λ being a characteristic time associated with the stable Lyapunov modes. In practice, we are interested in averages of quantities on the attractor $g(\mathbf{u}^{\mathcal{A}}(t))$, but we can only calculate quantities $g(\mathbf{u}(t))$ that will include some effect—if small—of the spin-up transient.

Next, we seek to quantify the effect of this gap on estimates $J_T \approx J_\infty$. Consider the computation of J_T. In (7), we have effectively found an estimate of

$$J_T^{\mathcal{A}} = \frac{1}{T_s} \int_{t_0}^{t_0+T_s} g(\mathbf{u}^{\mathcal{A}}(t)) dt, \quad (48)$$

by choosing t_0 sufficiently large. We now want to consider an error model of the form

$$e_{T,hp} = \underbrace{(J_{T,hp} - J_T)}_{e_{hp}} + \underbrace{(J_T - J_T^{\mathcal{A}})}_{e_\lambda} + \underbrace{(J_T^{\mathcal{A}} - J_\infty)}_{e_T}, \quad (49)$$

where a new error e_λ is introduced, associated with the spin-up transient. The model for e_T in (9) will apply without modification, while

the model for e_{hp} will be subject to slightly different assumptions, where in (8), C_q was bounded by the value on the attractor, $\mathcal{A}$, here we must assume that C_q is bounded from $t = 0$ to $t = t_0 + T_s$, including both the attractor *and* the transient part of the trajectory. We only require that the transient part be in the basin of attraction of $\mathcal{A}$, $B(\mathcal{A})$. We assume that a model of the form used in (8) applies in expectation when the transient component is included.

Next, we concentrate on e_λ,

$$J_T - J_T^{\mathcal{A}} = \int_{t_0}^{t_0+T_s} \left(g(\mathbf{u}(t)) - g(\mathbf{u}^{\mathcal{A}}(t))\right) dt. \quad (50)$$

We now assume that, like $\mathbf{u}$, g will decay exponentially in t as (47), such that

$$g(\mathbf{u}(t)) - g(\mathbf{u}^{\mathcal{A}}(t)) \equiv \delta g^{\mathcal{A}}(t) \approx A_\lambda \exp\left(-\frac{t}{T_\lambda}\right) \quad (51)$$

will apply for $t \in [0, \infty)$, with A_λ being a constant that can be related to the deviation between $g(\mathbf{u}(0))$ and $g(\mathbf{u}^{\mathcal{A}}(0))$.

From this assumption,

$$\begin{aligned} e_\lambda &= \frac{1}{T_s} \int_{t_0}^{t_0+T_s} g(\mathbf{u}(t)) - g(\mathbf{u}^{\mathcal{A}}(t)) dt \\ &\approx \frac{1}{T_s} \int_{t_0}^{t_0+T_s} A_\lambda \exp\left(-\frac{t}{T_\lambda}\right) dt \\ &= A_\lambda \frac{T_\lambda}{T_s} \exp\left(-\frac{t_0}{T_\lambda}\right) \left(1 - \exp\left(-\frac{T_s}{T_\lambda}\right)\right). \end{aligned} \quad (52)$$

Taking the absolute value, we can find a bounding model

$$|e_\lambda| = |A_\lambda| \frac{T_\lambda}{T_s} \exp\left(-\frac{t_0}{T_\lambda}\right). \quad (53)$$

As before, manipulation of (49) allows

$$|e_{T,hp}| = |e_{hp} + e_\lambda + e_T| \quad (54)$$
$$\leq |e_{hp}| + |e_\lambda| + |e_T|. \quad (55)$$

Now, we take an expectation of the absolute value of $e_{T,hp}$,

$$\mathbb{E}[|e_{T,hp}|] \leq \mathbb{E}_{B(\mathcal{A})}[|e_{hp}|] + \mathbb{E}_{\text{IC}}[|e_\lambda|] + \mathbb{E}_{\mathcal{A}}[|e_T|], \quad (56)$$

where $\mathbb{E}_{B(\mathcal{A})}$ gives the expectation on the basin of attraction of $\mathcal{A}$. Here, the expectation of $|e_{T,hp}|$ does not reduce to an expectation *on the attractor*. The statistical term is handled on the attractor as before, and we have assumed that the discretization error is bounded by the same form in expectation on $B(\mathcal{A})$ as on $\mathcal{A}$. Finally, the expectation of $|e_\lambda|$ is taken on the set of initial conditions used. This allows us to take the expectation of (53) to complete (56). Because we anticipate T_λ will be bounded by a constant for a given system, this is given by

$$\mathbb{E}_{\text{IC}}[|e_\lambda|] = \mathbb{E}_{\text{IC}}[|A_\lambda|] \frac{T_\lambda}{T_s} \exp\left(-\frac{t_0}{T_\lambda}\right). \quad (57)$$

If a A_λ and T_λ can be identified by observation of $g(\mathbf{u}(t))$ given an initial condition $\mathbf{u}_{\text{IC}}$, $|e_\lambda|$ is no longer stochastic and the $\mathbb{E}[|e_{T,hp}|] \to |e_{T,hp}|$ as in (53).

Putting all the pieces together, we can now give an error model that incorporates the effects of spin-up and ensemble estimation

$$e_{\text{model,MC}} = \tilde{A}_\lambda \frac{T_\lambda}{T_{s,\text{MC}}} \exp\left(-\frac{t_0}{T_\lambda}\right) + C_q (\Delta t)_{\text{MC}}^q + \frac{A_0}{\sqrt{M_{\text{ens}}}} T_{s,\text{MC}}^{-r},$$

$$(58)$$

where $\tilde{A}_\lambda$ can be either estimated on an instance-by-instance basis or by estimating the expectation on the family of initial conditions. Under this model, e_λ will scale with the exponent of a large negative value when $t_0 \gg T_\lambda$. Even when t_0/T_λ, (53) suggests that the decay-induced error term will still scale with T_s^{-1}, faster than the expected CLT rate of $T_s^{-1/2}$, and thus, it will be dominated it as $T_s \gg 1$. This implies two "paths" to control spin-up errors: either choosing t_0 long enough to shrink the mean offset error from t_0 or choosing T_s long enough so that the mean offset contribution to the simulation error is small in spite of the error at $t = t_0$.

C. Identification of the spin-up transient model

We will now develop a method to fit the error model. In order to do so, consider observations $g_n \equiv g(t_n)$ and $g_n^A \equiv g^A(t_n)$ for t_n in $\{t_0, t_0 + N_{\text{skip}}\Delta t, \ldots, t_0 + T_s\}$. We will assume that N_{skip} is large enough that the solution at each t_n is effectively independent. If this is the case, then we can assume that each g_n^A will be an independent and identically distributed (i.i.d.) draw from a bounded, stationary distribution with mean J_∞. The distributions of $g^A(\mathbf{u}(t))$ and $g(\mathbf{u}(t))$, in general, are not known. In order to facilitate an estimate of the mean behavior, we will assume g_n^A are i.i.d. drawn from a normal distribution with mean value J_∞. Then, we have

$$g_n \sim \mathcal{N}(J_\infty + \delta g^A(t_n), \sigma_g^2), \qquad (59)$$

where the relationship between g_n and g_n^A is taken from (51).

In order to understand the implications of this model, we can use a set of reference RK4 simulations of the Lorenz system with $N_t = 10^5$ timesteps sampled without spin-up over a period $T = 100$ from initial conditions similar to those given in (22) with a scaled-up standard deviation of 100 in all three variables to highlight the initial transient. In order to treat each of J_∞, σ_g, A_λ, and T_λ in (59) as unknowns, we use Hamiltonian Monte Carlo with the likelihood function implied in (59). We discard from $t = 0$ to $t = 5$ and then take 10 000 equispaced samples from $t = 5$ to $t = 100$. For prior models, we start by computing naïve estimators of the mean and standard deviation of the trace, $\tilde{J}$ and $\tilde{\sigma}$ using the downsampled trace signal $\{g_n\}$ and then use

$$\begin{aligned} J_\infty &\sim \mathcal{N}(\tilde{J}, \tilde{\sigma}^2), \\ \sigma_g &\sim \Gamma(\alpha_\sigma, \beta_\sigma), \\ A_\lambda &\sim \mathcal{N}(0, \max(g_h p) - \min(g_h p)), \\ T_\lambda &\sim \Gamma(\alpha_T, \beta_T), \end{aligned} \qquad (60)$$

where

$$\begin{aligned} (\alpha_\sigma, \beta_\sigma) &\Leftarrow \left(\mu_\sigma = \tilde{\sigma}, \sigma_\sigma = \frac{\tilde{\sigma}}{10}\right), \\ (\alpha_T, \beta_T) &\Leftarrow (\mu_T = 10.0, \sigma_T = 10.0). \end{aligned}$$

It should be noted that in this specification, the Bayesian fit only requires a user-supplied prior for the decay time and for the uncertainty in the standard deviation, assumptions upon which the fitting method only requires to be reasonable.
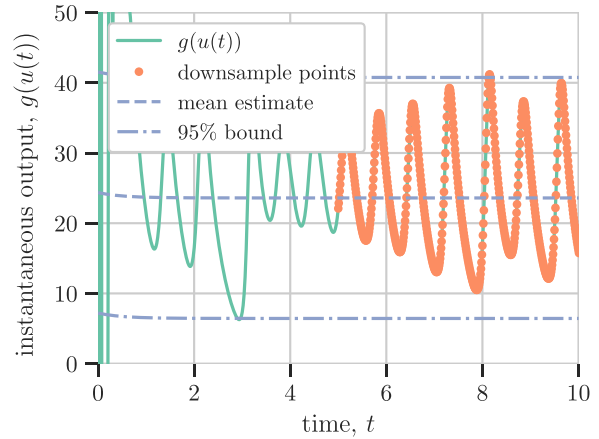

FIG. 13. $g = u_2(t)$ trace in the transient region with Bayesian method fit.

A sample fit and trace are found in Fig. 13, for which the maximum *a posteriori* (MAP) estimate gives $T_\lambda = 0.312$ and $A_\lambda = -0.925$. For the Lorenz system, the initial transient onto the attractor is very rapid, almost negligible. Applying the Bayesian fit procedure to an ensemble of 1000 runs generated in the same way as Fig. 13, we can find maximum *a posteriori* (MAP) estimates of the variables T_λ and $|A_\lambda|$ in the decay model. In Figs. 14 and 15, histograms of these variables are shown, which are needed to determine (53). We can see that the fit procedure identifies values

$$\begin{aligned} T_\lambda &< 4.03, \\ |A_\lambda| &< 38.7, \end{aligned} \qquad (61)$$

for greater than 97% of initial conditions, up to two standard deviations above the mean. Using these values as a conservative estimate for the mean offset, we can now model the effect of the transient behavior.
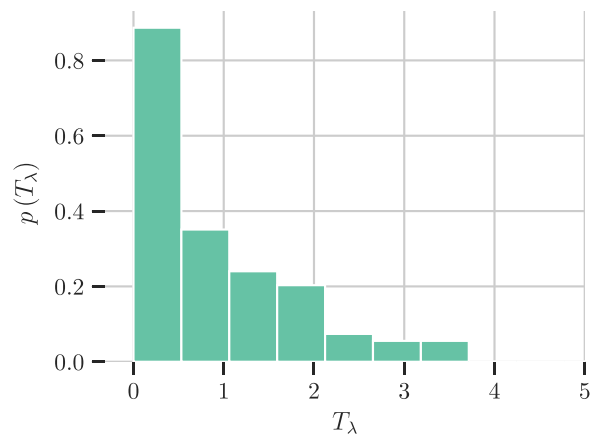

FIG. 14. MAP estimate T_λ for Lorenz system transient. Collected over 1000 Lorenz trajectories with $\Delta t = 10^{-2}$, $T_s = 100$, and randomized $\mathbf{u}_{\text{IC}}$. Outliers truncated, greater than 97% of data in the pictured range.
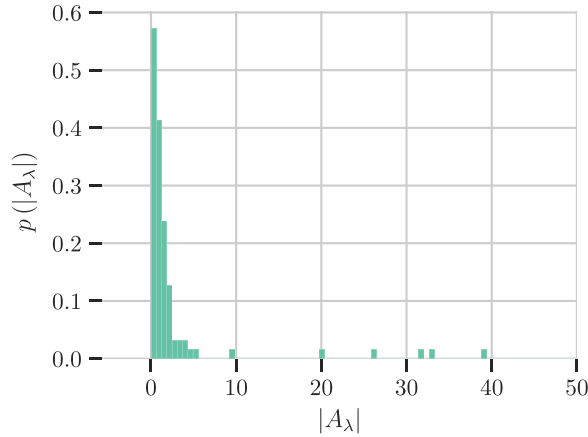
FIG. 15. MAP estimate $|A_\lambda|$ for Lorenz system transient. Collected over 1000 Lorenz trajectories with $\Delta t = 10^{-2}$, $T_s = 100$, and randomized $\mathbf{u}_{\text{IC}}$. Outliers truncated, greater than 97% of data in the pictured range.

FIG. 16. Dependence of normalized error expectation $e_{\text{model,MC}}/\sigma_g$ on normalized time step $\Delta t/T_d$ and normalized spin-up time t_0/T_d with total cost set at $U = 1.2 \times 10^6$ for forward Euler. Red star denotes optimum, and dashed line indicates optimal t_0 given Δt.

VII. OPTIMAL TIME-STEPPING INCLUDING SPIN-UP

Now we can consider how the cost and error impact of spin-up is incorporated into the model for error at a fixed cost. The spin-up time requires the use of N_0 timesteps

$$N_0 = \left\lceil \frac{t_0}{(\Delta t)_{\text{MC}}} \right\rceil \approx \frac{t_0}{(\Delta t)_{\text{MC}}},$$ (62)

with N being the total number of timesteps used, given by

$$N = N_0 + N_{\text{MC}} = \frac{t_0}{(\Delta t)_{\text{MC}}} + N_{\text{MC}},$$ (63)

where N_{MC} is the number of timesteps during sampling for t_0 to $t_0 + T_s$ on a given instance.

By normalizing (58) then substituting (63), we arrive at a transient-inclusive non-dimensional model for the error

$$\left(\frac{e_{\text{model}}}{\sigma_g}\right)_{\text{MC}} = \frac{\tilde{A}_\lambda}{\sigma_g}\frac{T_\lambda}{T_d}\left(N\left(\frac{\Delta t}{T_d}\right)_{\text{MC}} - \frac{t_0}{T_d}\right)^{-1}\exp\left(-\frac{t_0/T_d}{T_\lambda/T_d}\right)$$
$$+ \frac{C_q T_d^q}{\sigma_g}\left(\frac{\Delta t}{T_d}\right)_{\text{MC}}^q + \sqrt{\frac{2}{\pi}}M_{\text{ens}}^{-\frac{1}{2}}\left(N\left(\frac{\Delta t}{T_d}\right)_{\text{MC}} - \frac{t_0}{T_d}\right)^{-\frac{1}{2}}.$$
(64)

Using this result, we can solve numerically for $(\Delta t)_{\text{MC,opt}}$ and $e_{\text{MC,opt}}$ via (64).

Consider a Lorenz simulation on which a budget of $U = pN = 1.2 \times 10^6$ right-hand side evaluations are available on each of M_{ens} parallel processors. We start by studying the error under (64) as Δt and t_0 vary with a conservative estimate for the transient behavior using the bounding values in (61). In Fig. 16, we show e_{model} for forward Euler at a fixed cost of $U = 1.2 \times 10^6$. (The optimum is denoted by a red star.) Moving to the right, discretization error becomes the dominant factor as $\Delta t \gg T_d$. The diagonal boundary gives the region of feasibility at which, under the cost constraint, sampling no longer occurs ($T_s = 0$). Moving from the optimum toward the bottom left,
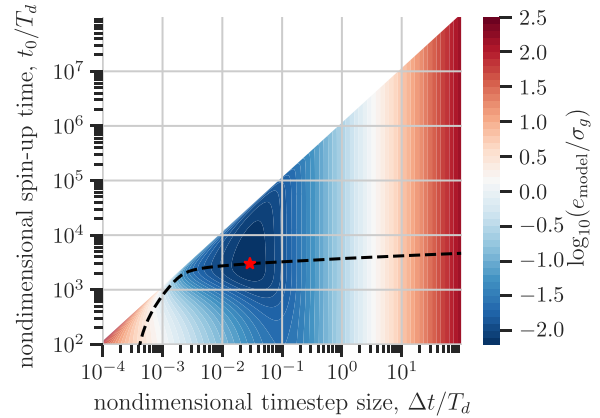
$t_0 \rightarrow 0$, $T_s \rightarrow 0$, and $\Delta t \ll T_d$; thus the transient error and sampling error become dominant. Similar plots for RK3 and RK4 are found in Figs. 17 and 18. The optimal errors and optimizing simulations are described in Table V. We can see from these results that, at a fixed budget with $U = 1.2 \times 10^6$, the effect of increasing the discretization order makes a smaller error possible with a larger time step, which means fewer timesteps to traverse the spin-up time. These two effects combine to allow for an increase in the sampling time available T_s, allowing significantly less sampling error for RK3 compared to FE, and an additional—albeit smaller—benefit moving from RK3 to RK4, holding cost fixed.

In Fig. 19, we take another perspective on these results for RK3 by varying Δt and plotting the optimal t_0, T_s, and e_{model}. As Δt gets large, the optimal choice of t_0 has logarithmic growth, and when $\Delta t/T_d \ll 1$, the optimal choice of t_0 rapidly falls to zero.
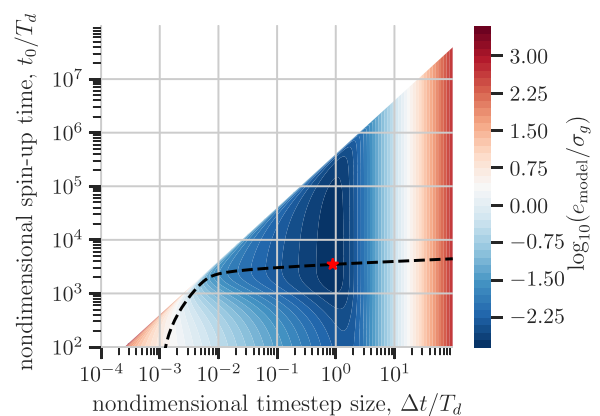

FIG. 17. Dependence of normalized error expectation $e_{\text{model,MC}}/\sigma_g$ on normalized time step $\Delta t/T_d$ and normalized spin-up time t_0/T_d with total cost set at $U = 1.2 \times 10^6$ for third-order Runge–Kutta. Red star denotes optimum, and dashed line indicates optimal t_0 given Δt.
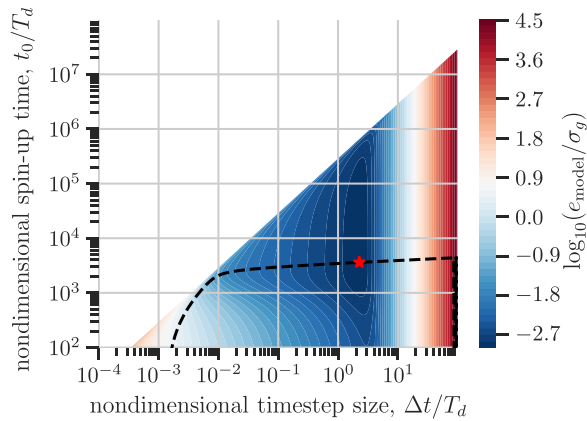
FIG. 18. Dependence of normalized error expectation $e_{model,MC}/\sigma_g$ on normalized time step $\Delta t/T_d$ and normalized spin-up time t_0/T_d with total cost set at $U = 1.2 \times 10^6$ for fourth-order Runge–Kutta. Red star denotes optimum, and dashed line indicates optimal t_0 given Δt.

Parallelization has a small but non-zero effect on the optimal choice of sample time. The sampling time also has a small effect from parallelization, in this case constrained to a small region. Outside that Δt region, T_s scales with Δt both as $\Delta t \rightarrow 0$ and as $\Delta t \rightarrow \infty$.

The bottom plot of Fig. 19 shows the variation of error with Δt. In this plot, we can see three distinct regions. For $\Delta t/T_d \gg 10^{-2}$, discretization error is the dominating error, and the convergence goes with the discretization error rate. Approaching the optimum, sampling error becomes the dominant error contribution, starting at $\Delta t \approx 2 \times 10^{-2}$ until $\Delta t \approx 10^{-3}$. In this region, the convergence is around the CLT-implied $1/2$ rate, and the effect of parallelization is clearly seen. For $\Delta t \lesssim 10^{-3}$, however, the spin-up error becomes the dominant error contribution. The optimal choice of t_0 begins to fall rapidly, as the sampling and spin-up must compete for computational resources under the budget. Once the spin-up error dominates, the paradigm by which (53) is controlled shifts from the $\exp(-t_0)$ term to the T_s^{-1} term as $\Delta t/T_d \rightarrow 0$, since resolving T_s delivers both spin-up and sampling error control.

This interdependence will evidently have an effect on the overall scaling between cost and error, which we now seek to understand. Here, we study the variation of $e_{model,MC}$ with U under the optimal choices and evaluate how well $e_{model,MC}$ approximates experimental data for $\mathbb{E}[|J_{MC} - J_\infty|]$. In Fig. 20, the variation of $e_{model,MC}$ computed via (64) as a function of M_{ens} and U is shown. From this figure, we can see that, in the limit of small error, the sampling costs dominate and the best possible rate is given by the estimate in (44), limited by the CLT. On the other hand, when the cost is more moderate, scaling of

TABLE V. Optimal Lorenz simulations for output $g = u_2$ under budget of $U = 1.2 \times 10^6$ right-hand side evaluations using $M_{ens} = 1$.

Method	p	e_{model}	Δt	t_0	T_s
FE	1	0.0502	2.54×10^{-4}	30.2	275
RK3	3	0.0130	8.52×10^{-3}	35.5	3370
RK4	4	8.89×10^{-3}	0.0224	36.7	6670

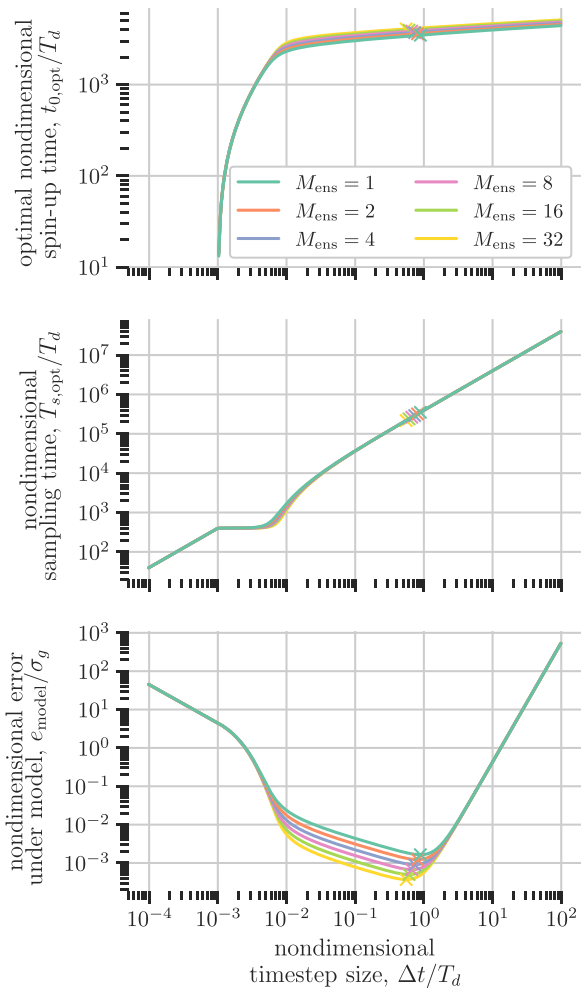**FIG. 19.** Dependence of normalized spin-up time t_0/T_d, sampling time T_s/T_d, and model error e_{model}/σ_g on normalized time step $\Delta t/T_d$ with total cost set at $U = 1.2 \times 10^6$ for third-order Runge–Kutta.

the error is close to the discretization error convergence rate in (19). In this region, the spin-up costs are significant, and high-order discretization brings the state more efficiently to the start of sampling. In the spin-up dominated region, the effect of the parallel ensemble approach is minimal since spin-up must be overcome on each processor.

Now, we validate the total error model for the Lorenz system by a final numerical experiment. At each choice of M_{ens} and U, we generate 1000 individual realizations of J_{MC} at the computed $(\Delta t)_{MC,opt}$ and $N_{MC,opt}$ and using the model fit given in Table II. In Figs. 21–23, we show the predictions and the results of Monte Carlo estimates of $\mathbb{E}[|J_{MC} - J_\infty|]$ for our three discretizations. These results validate the model with significant discrepancies only when the asymptotic assumptions—Δt small and T_s large—do not hold, due to budget limitations in the limit of small U.

VIII. CONCLUSIONS AND FORTHCOMING WORK

In this manuscript, we have developed a theoretical framework for the total error incurred by the discrete sampling of mean outputs
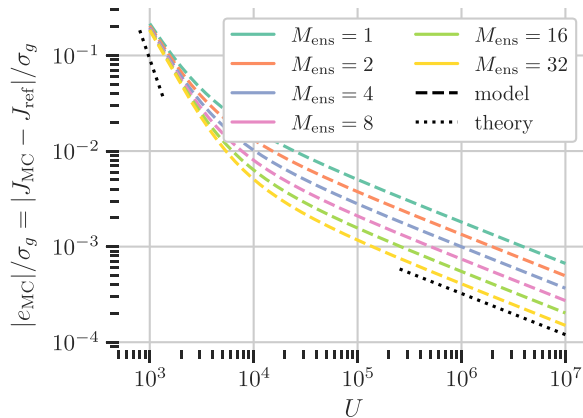
FIG. 20. Optimal non-dimensional error under model as a function of total cost U for RK3. Theory totem on the left-hand side: discrete convergence rate, $1/q$; on right-hand side: $\frac{2(q+r)}{q}$ rate from (44).
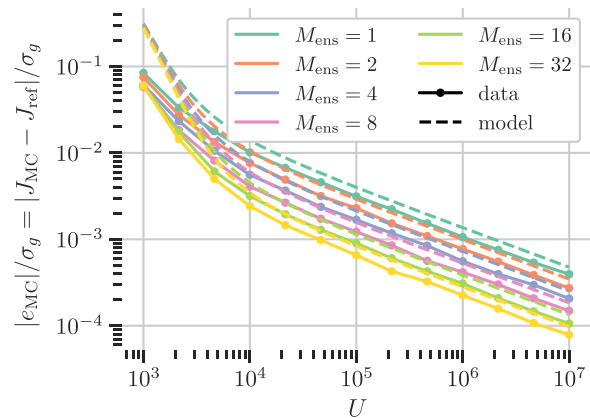

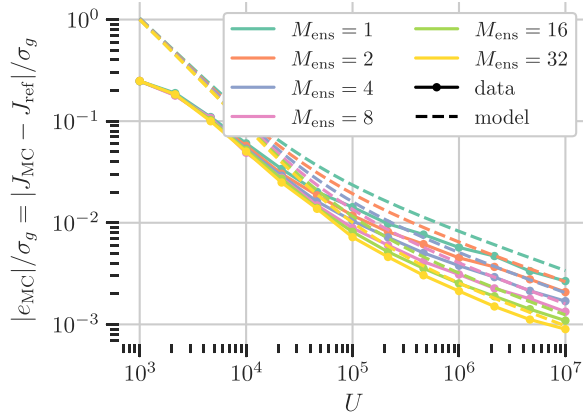**FIG. 23.** Total cost model and Monte Carlo validation as a function of total cost U for RK4.

FIG. 21. Total cost model and Monte Carlo validation as a function of total cost U for FE.
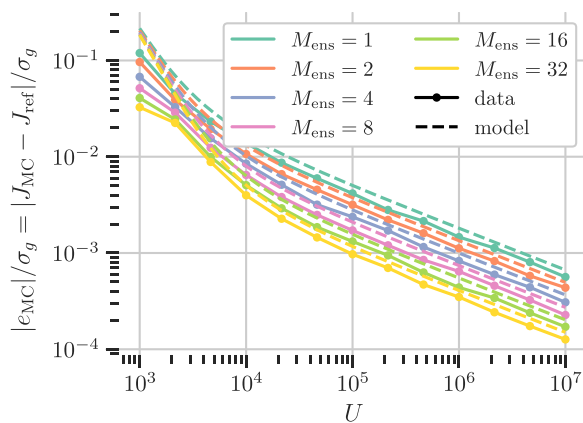

FIG. 22. Total cost model and Monte Carlo validation as a function of total cost U for RK3.

of ergodic ODEs. These findings are validated by Monte Carlo studies of the Lorenz system using Runge–Kutta methods. We incorporate effects of parallelization and spin-up and validate that the models match observed results in experiments. Using these models, we are able to develop a comprehensive understanding of the relationship between the wall-clock cost of a simulation and the amount of error in expectation that it might achieve.

The first result of these studies is to demonstrate the central limit theorem that gives a lower limit of $-1/2$ on the convergence rate of error with respect to sampling cost in a simulation. The effect of discretization error is to penalize the observed convergence rate away from this limiting convergence rate, causing a higher convergence rate than the $-1/2$ rate for the central limit theorem. In order to minimize the error that might be expected in a given simulation, there exists a specific choice of $T_s = N_s \Delta t$ that achieves the minimum in expectation given a number of timesteps N_s available for computation.

In addition to this finding, we show that ensemble simulations reduce the sampling error, albeit with a similar penalty for discrete simulations. Moreover, the effect of spin-up costs—assuming optimal choices—is to increase the error at a given amount of the total simulation cost, which cannot be reduced by parallelization. The effect of spin-up, however, is negligible in the limit of infinite costs, but the results suggest it should be considered in cost-constrained contexts.

A key problem with the applicability of this research presented in this paper is the expense of identifying the parameters of the error model. In order to overcome this, we believe that leveraging a Bayesian approach as in Oliver et al.[20] can allow us to approximate the model in (20) at relatively small cost, and then exploit the result to conduct a high-fidelity simulation at (approximately) optimal discretizations. Furthermore, the framework must be extended to handle chaotic partial differential equation systems as opposed to ordinary differential equation systems. Though many discrete partial differential equation (PDE) systems are discretized in a form that reduces to an ODE system, a rigorous model for the error and cost of a PDE system should account for the contributions of both temporal discretization *and* spatial discretization. These will be the primary concerns of our forthcoming work. Finally, for the most expensive systems, it is likely

that cost constraints will make simulations that reside comfortably inside the asymptotic convergence region prohibitively expensive. Further investigation of the convergence behavior as Δt leaves the asymptotic region or at the convergence limit is left to future research.

ACKNOWLEDGMENTS

The authors would like to acknowledge the support of The Boeing Company (technical monitor Dr. Andrew Cary).

AUTHOR DECLARATIONS
Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Cory V. Frontin: Conceptualization (equal); Data curation (equal); Formal analysis (equal); Software (equal); Writing – original draft (equal). **David L. Darmofal:** Conceptualization (equal); Formal analysis (equal); Funding acquisition (equal); Project administration (equal); Resources (equal); Software (equal); Supervision (equal); Writing – review & editing (equal).

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1]S. T. Bose and G. I. Park, "Wall-modeled large-eddy simulation for complex turbulent flows," Annu. Rev. Fluid Mech. **50**, 535 (2018).
[2]A. Garai, L. T. Diosady, S. M. Murman, and N. K. Madavan, "Scale-resolving simulations of bypass transition in a high-pressure turbine cascade using a spectral element discontinuous Galerkin method," J. Turbomach. **140**, 031004 (2018).
[3]R. Stoll, J. A. Gibbs, S. T. Salesky, W. Anderson, and M. Calaf, "Large-eddy simulation of the atmospheric boundary layer," Boundary-Layer Meteorol. **177**, 541–581 (2020).
[4]K. Duraisamy, "Perspectives on machine learning-augmented Reynolds-averaged and large eddy simulation models of turbulence," Phys. Rev. Fluids **6**, 050504 (2021).
[5]L. T. Diosady and S. M. Murman, "Higher-order methods for compressible turbulent flows using entropy variables," in *53rd AIAA Aerospace Sciences Meeting* (AIAA, 2015), p. 0294.
[6]M. Franciolini and S. M. Murman, "Multigrid preconditioning for a space-time spectral-element discontinuous-Galerkin solver," in *AIAA Scitech 2020 Forum* (AIAA, 2020), p. 1314.
[7]M. J. Lighthill, "The recently recognized failure of predictability in Newtonian dynamics," Proc. R. Soc. London, Ser. A **407**, 35–50 (1986).
[8]J.-P. Eckmann and D. Ruelle, "Ergodic theory of chaos and strange attractors," in *The Theory of Chaotic Attractors* (Springer, 1985), pp. 273–312.
[9]D. R. Chapman, "Computational aerodynamics development and outlook," AIAA J. **17**, 1293–1313 (1979).
[10]P. Spalart, W. Jou, M. Strelets, S. Allmaras *et al.*, "Comments on the feasibility of LES for wings, and on a hybrid RANS/LES approach," in Advances in DNS/LES, 1997.

[11]H. Choi and P. Moin, "Grid-point requirements for large eddy simulation: Chapman's estimates revisited," Phys. Fluids **24**, 011702 (2012).
[12]X. I. Yang and K. P. Griffin, "Grid-point and time-step requirements for direct numerical simulation and large-eddy simulation," Phys. Fluids **33**, 015108 (2021).
[13]J. Kim, P. Moin, and R. Moser, "Turbulence statistics in fully developed channel flow at low Reynolds number," J. Fluid Mech. **177**, 133–166 (1987).
[14]A. Lozano-Durán and J. Jiménez, "Effect of the computational domain on direct simulations of turbulent channels up to $Re_\tau = 4200$," Phys. Fluids **26**, 011702 (2014).
[15]J. C. Del Álamo, J. Jiménez, P. Zandonade, and R. D. Moser, "Scaling of the energy spectra of turbulent channels," J. Fluid Mech. **500**, 135–144 (2004).
[16]K. A. Goc, O. Lehmkuhl, G. I. Park, S. T. Bose, and P. Moin, "Large eddy simulation of aircraft at affordable cost: A milestone in computational fluid dynamics," Flow **1**, E14 (2021).
[17]R. L. Thompson, L. E. B. Sampaio, F. A. de Bragança Alves, L. Thais, and G. Mompean, "A methodology to evaluate statistical errors in DNS data of plane channel flows," Comput. Fluids **130**, 1–7 (2016).
[18]S. Russo and P. Luchini, "A fast algorithm for the estimation of statistical error in DNS (or experimental) time averages," J. Comput. Phys. **347**, 328–340 (2017).
[19]C. Mockett, T. Knacke, and F. Thiele, "Detection of initial transient and estimation of statistical error in time-resolved turbulent flow data," in *Proceedings of the 8th International Symposium on Engineering Turbulence Modelling and Measurements* (European Research Collaboration on Flow Turbulence and Combustion, 2010), pp. 9–11.
[20]T. A. Oliver, N. Malaya, R. Ulerich, and R. D. Moser, "Estimating uncertainties in statistics computed from direct numerical simulation," Phys. Fluids **26**, 035101 (2014).
[21]M. Lee and R. D. Moser, "Direct numerical simulation of turbulent channel flow up to $Re_\tau \approx 5200$," J. Fluid Mech. **774**, 395–415 (2015).
[22]A. M. Stuart, "Numerical analysis of dynamical systems," Acta Numer. **3**, 467–572 (1994).
[23]M. Denker, "The central limit theorem for dynamical systems," Banach Cent. Publ. **23**, 33–62 (1989).
[24]R. C. Bradley, "Basic properties of strong mixing conditions. a survey and some open questions," Probab. Surv. **2**, 107–144 (2005).
[25]V. Araújo, I. Melbourne, and P. Varandas, "Rapid mixing for the Lorenz attractor and statistical limit laws for their time-1 maps," Commun. Math. Phys. **340**, 901–938 (2015).
[26]E. Hairer, *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems*, 2nd ed., Springer Series in Computational Mathematics Vol. 14 (Springer, Berlin, Germany, 1993).
[27]D. Viswanath, "Global errors of numerical ODE solvers and Lyapunov's theory of stability," IMA J. Numer. Anal. **21**, 387–406 (2001).
[28]E. N. Lorenz, "Deterministic nonperiodic flow," J. Atmos. Sci. **20**, 130–141 (1963).
[29]C. Sparrow, *The Lorenz Equations: Bifurcations, Chaos, and Strange Attractors* (Springer Science & Business Media, 1982).
[30]J. Dormand, R. Duckers, and P. Prince, "Global error estimation with Runge-Kutta methods," IMA J. Numer. Anal. **4**, 169–184 (1984).
[31]K. E. Trenberth, "Some effects of finite sample size and persistence on meteorological statistics. Part I: Autocorrelations," Mon. Weather Rev. **112**, 2359–2368 (1984).
[32]F. Harris, "On then use of windows for harmonic analysis with the discrete Fourier transform," Proc. IEEE **66**, 51 (1978).
[33]G. Karniadakis and S. Sherwin, *Spectral/hp-Element Methods for Computational Fluid Dynamics* (Oxford University Press, 2005).
[34]V. Makarashvili, E. Merzari, A. Obabko, A. Siegel, and P. Fischer, "A performance analysis of ensemble averaging for high fidelity turbulence simulations at the strong scaling limit," Comput. Phys. Commun. **219**, 236–245 (2017).