Output error behavior for discretizations of ergodic, chaotic ODE systems

Cory V. Frontin\textsuperscript{1,\textit{a}} and David L. Darmofal\textsuperscript{1,\textit{b}}

Department of Aeronautics and Astronautics, Massachusetts Institute of Technology

(Dated: 17 October 2022)

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 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

For chaotic systems, estimation of long-time behavior is challenging because chaotic systems have limited predictability. Of the general class of chaotic systems, a subset are ergodic systems, whose long-term states are drawn from a stationary distribution, independent of initial condition. For ergodic chaotic problems, we frequently want to quantify the unique infinite-time average of some instantaneous quantity of interest of the system:

\[ J_\infty = \lim_{T \to \infty} \frac{1}{T} \int_0^T g(u(t)) \, dt, \]  

where \( g \) is the instantaneous output functional, and the state \( u(t) \) is governed by a dynamical system of the form:

\[ \frac{du}{dt} = f(u) \]  

with a given initial condition (IC), \( u(0) = u_{IC} \).

Often, the complexity of a chaotic systems of interest is high, and accordingly the cost of an accurate computational estimate of \( J_\infty \) becomes formidable. As the cost of computational simulation gets larger, efficient discretization methods become critical for accurately estimating quantities of interest.

Understanding the error in approximations of \( J_\infty \) is nontrivial because statistical errors (errors due to finite-time approximation) and discretization error (error due to 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 sampling time at some large number of characteristic times and validate that discretization error converges as expected, assuming negligible sampling error. Recent work has sought to quantify the effect of statistical error more robustly, using turbulent flow theory, advanced spatio-temporal statistical post-processing methods, statistical windowing techniques, or by extending the concept of Richardson extrapolation to chaotic flows using auto-regressive models and Bayesian methods. The latter work is notable for its use to estimate the statistical errors in the DNS of a high-Re turbulent channel flow.

The objective of this paper is to investigate the behavior of statistical and discretization errors as a function of computational cost for ergodic systems. Following a similar approach to Oliver et al., 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 timestep size decreases. However, it does not increase exponentially with sampling time as might be expected from classical numerical analysis but rather asymptotes to a constant value with respect to sampling time. Further, for a given computational cost (e.g. number of timesteps), an optimal choice of discretization (i.e. timestep) exists that minimizes the expected error in a 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 $J_{\infty}$, we compute finite-time, discrete estimates of the outputs of interest of the true system:

$$J_{T,hp} = \frac{1}{T_s} \int_{t_0}^{t_0+T_s} \left( g_{hp}(u_{hp}(t)) \right) dt,$$

where the notation $I_a^b(\cdot)$ here represents the quadrature approximation of the integral $\int_a^b(\cdot) \, dt$ 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:

$$\max_{t \in [0,t_0+T_s]} |g_{hp}(u_{hp}(t)) - g(u(t))| = O(h^p)$$

when the discretization is applied to a well-posed (non-chaotic) system. Then we sample that 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

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

By introducing a third value,

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

3
we can re-write the error using an identity:

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

Here, we define the “discretization error” and “sampling error”, respectively:

\[ e_{hp} \equiv J_{T,hp} - J_T \] (8)
\[ e_T \equiv J_T - J_\infty. \] (9)

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

\[ |e_{T,hp}| = |e_{hp} + e_T| \leq |e_{hp}| + |e_T|. \] (10)

Thus, the total error incurred by approximation is bounded by the sum of the absolute discretization and sampling errors. Next, we define the attractor of the operator \( f, A \), as the set of long-term states towards which all trajectories converge independently of initial condition\(^{15} \). We can define the expectation \( \mathbb{E}_A[\phi(u_0)] \) for a generic function \( \phi \) as the expectation taken over all the trajectories that can result from starting from points on the attractor, \( A \):

\[ \mathbb{E}_A[\phi] = \frac{1}{|A|} \int_{u_0 \in A} \phi(u_0) \, du_0. \] (11)

For the case in question we will be considering either

\[ \phi(u_0) = \frac{1}{T_s} \left| \int_{t_0}^{t_0+T_s} g(u(t)) \, dt - J_\infty \right|, \]

or

\[ \phi(u_0) = \frac{1}{T_s} \left| \int_{t_0}^{t_0} \left( g_{hp}(u_{hp}(t)) \right) - \int_{t_0}^{t_0+T_s} g(u(t)) \, dt \right|, \]

with, for these examples, \( u(t_0) = u_0 \in A \). Given these definitions, we can now take the expectation of (10), giving

\[ \mathbb{E}_A[|e_{T,hp}|] \leq \mathbb{E}_A[|e_{hp}|] + \mathbb{E}_A[|e_T|] \] (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):

\[ e_T = \frac{1}{T_s} \int_{t_0}^{t_0+T_s} g(u(t)) \, dt - J_\infty. \] (13)
Assuming that we choose $t_0$ such that each $u_0$ is effectively an independent sample from the attractor’s stationary distribution, then the quantity $g(u(t))$ 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 $\alpha$-mixing is met\cite{16,17}, which has been shown for the Lorenz system\cite{18}. Thus we can write $e_T$ as:

$$\frac{1}{T_s} \approx N\left(0, \left(\frac{\pi}{2} A_0 T_s^{-1/2}\right)^2\right),$$  \hspace{1cm} (14)

where $N(\mu, \sigma^2)$ gives the normal distribution with mean $\mu$ and variance $\sigma^2$. If we take the absolute value of this random variable, the result is a halfnormal distribution:

$$|e_T| \approx H\left(\left(\frac{\pi}{2} A_0 T_s^{-1/2}\right)^2\right),$$  \hspace{1cm} (15)

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

$$\mathbb{E}_A[|e_T|] \approx A_0 T_s^{-1/2}$$  \hspace{1cm} (16)

as $T_s$ goes to infinity.

Now consider the use of a time-stepping method to give a discrete approximation $u_{hp}(t_n)$ of $u(t_n)$ for each $t_n = n(\Delta t)$. Following classical analysis\cite{19}, we might expect that the discretization error should take a form:

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

This analysis is based on bounding the growth of local truncation error at each timestep by the Lipschitz constant, $\Lambda$, of the underlying system, with $C_p$ a constant parameter that depends on the choice of method. However, Viswanath showed\cite{20} that, the global error could be modeled by a form:

$$|e_{hp}| \approx E(T_s; p)(\Delta t)^p,$$  \hspace{1cm} (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:

$$\mathbb{E}_A[|e_{hp}|] \approx C_p (\Delta t)^p,$$  \hspace{1cm} (19)
As our results in Section 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:

\[
\mathbb{E}_A[|e_{T,hp}|] \leq e_{\text{model}} = C_q(\Delta t)^q + A_0 T_s^{-r},
\]

that bounds \(\mathbb{E}_A[|e_{T,hp}|]\) when \(\Delta t\) is small enough and \(T_s\) is large enough to satisfy the asymptotic assumptions. Here, \(q\) is the observed discretization convergence rate, which in practice may differ from \(p\) due to numerical cancellations 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 PROPOSED ERROR MODEL ON THE LORENZ SYSTEM

In the following section, 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

\[
\frac{du}{dt} = f(u; \alpha) = \begin{pmatrix} \alpha_0(u_1 - u_0) \\ u_0(\alpha_1 - u_2) - u_1 \\ u_0u_1 - \alpha_2u_2 \end{pmatrix},
\]

where \(u = [u_0, u_1, u_2]^T\) and \(\alpha = [\alpha_0, \alpha_1, \alpha_2]^T\). The Lorenz system is known to be chaotic for the classic Lorenz parametrization \cite{lorenz1963deterministic}. \(\alpha = [10, 28, 8/3]\), which is used everywhere in this text. For the output, we choose \(g(u) = u_2\). We consider a set of explicit methods: forward Euler (FE, \(p = 1\)), 3rd-order Runge-Kutta (RK3, \(p = 3\)), and 4th-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 \(O(\Delta t^p)\) for non-chaotic systems \cite{trefethen2000spectra}.

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:

\[
u_{\text{init}} \sim \begin{pmatrix} \mathcal{N}(1.0, 5.0^2) \\ \mathcal{N}(1.0, 5.0^2) \\ \mathcal{N}(1.0, 5.0^2) \end{pmatrix}.
\]

(22)
To guarantee that the initial sampling state \( u_0 \) 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 \( u_{\text{init}} \) for \( t_0 = 100 \) before proceeding to 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 \( J_\infty \) by a reference value \( J_{\text{ref}} \). \( J_{\text{ref}} \) 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 \( u_{\text{init}} \) as given in (22) and simulated using RK4 with \( \Delta t = 17.7 \times 10^{-6} \) and \( T_s = 6646.9 \). The resulting \( J_{\text{ref}} \) is:

\[
J_{\text{ref}} = 23.549916 \pm 0.000074, \quad (23)
\]

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

The computation of \( J_{\text{ref}} \) allows us to estimate errors \( e_{T, hp} \approx J_{T, hp} - J_{\text{ref}} \). For a given \( \Delta t, T_s \) pair, we then approximate \( \mathbb{E}_A[|e_{T, hp}|] \) 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 \( u_0^{(m)} \):

\[
\mathbb{E}[|e_{T, hp}|] \approx \frac{1}{M} \sum_{m=1}^{M} \left| J_{T, hp} \left( u_0^{(m)} \right) - J_{\text{ref}} \right|. \quad (24)
\]

In Figures 1, 2, and 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 \( \Delta t \) 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 eliminate non-convergent data at small \( T_s \) or large \( \Delta t \); 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 Figure 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
| method | $\Delta t_{\text{max}}$ | $T_{s,\text{min}}$ |
|--------|---------------------|------------------|
| FE     | $5.0 \times 10^{-3}$ | 1.0              |
| RK3    | $5.0 \times 10^{-2}$ | 1.0              |
| RK4    | $9.0 \times 10^{-2}$ | 1.0              |

**TABLE I: Fit boundaries for nonlinear least squares fits.**

|     | FE | RK3 | RK4 |     | FE | RK3 | RK4 |
|-----|----|-----|-----|-----|----|-----|-----|
| $A_0$ | 2.19 | 1.74 | 1.63 | $A_0$ | 1.94 | 1.50 | 1.41 |
| $r$   | 0.975 | 0.721 | 0.683 | $r$   | 0.820 | 0.648 | 0.620 |
| $C_q$ | 4995 | 942  | 85,900 | $C_q$ | 1410 | 1310 | 96,100 |
| $q$   | 1.65 | 2.70 | 4.83  | $q$   | 1.40  | 2.76 | 4.84  |

(a) $N_s = 10^4$

|     | FE | RK3 | RK4 |     | FE | RK3 | RK4 |
|-----|----|-----|-----|-----|----|-----|-----|
| $A_0$ | 1.52 | 0.978 | 0.918 | $A_0$ | 1.52 | 0.978 | 0.918 |
| $r$   | 0.693 | 0.553 | 0.538 | $r$   | 0.693 | 0.553 | 0.538 |
| $C_q$ | 714.6 | 2740 | 165,000 | $C_q$ | 714.6 | 2740 | 165,000 |
| $q$   | 1.273 | 2.96 | 5.02  | $q$   | 1.273 | 2.96 | 5.02  |

(b) $N_s = 10^5$

(c) $N_s = 10^6$

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

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 $\Delta t \to 0$.

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 timestep 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
FIG. 1: Expected relative error as a function of $\Delta t$ for Forward Euler discretization of the Lorenz equations. Nonlinear least squares fit based on $N_s = 10^6$ data.

Euler has $p = 1$ right-hand side evaluations), and thus $U_s = pN_s$. In Figure 5, we can see the effect of changing $\Delta t$ 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 3rd-order to 4th-order Runge-Kutta schemes is a only factor of about 1.4. Moreover, the results show that to achieve the lowest possible error, the optimal timestep will be discretization dependent. We investigate this further in the next section.
FIG. 2: Expected relative error as a function of $\Delta t$ for RK3 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 $\Delta t$ for RK4 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 $\Delta t$ for discretizations of the Lorenz equations.
FIG. 5: Expected percent error as a function of $\Delta t$ for discretizations of the Lorenz equations at a number of sampling residual evaluations. All fits evaluated at $U_s = 1.2 \times 10^6$. 

$E \left| e_{T_{hp}} \right| / \left| J_{ref} \right|$
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 error model in which the error is normalized by the standard deviation of the instantaneous output \( \sigma_g \) and the timescales \( \Delta t \) 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:

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

(25)

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

\[
A_0 = \sqrt{\frac{2}{\pi}} \sigma_g T_d^{1/2}. 
\]  

(26)

In general, \( T_d \) is hard to estimate accurately; this is a crux of the work of Oliver et al.\(^{13}\). In our formulation of the error model, we identify \( A_0 \), which avoids outright estimation of \( T_d \). However, for the purposes of understanding the behavior of the error, \( T_d \) is an intrinsic timescale which can be used to normalize \( \Delta t \) and \( T_s \).

The resulting non-dimensional form of the error model is

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

(27)

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

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

\[
\left( \frac{T_s}{T_d} \right)_{\text{opt}} = \left( \frac{1}{2\pi} \right)^{\frac{1}{2q+1}} \left( \frac{q C_q T_d^q}{\sigma_g} \right)^{-\frac{2}{2q+1}} N_s^{2q+1} 
\]  

(28)

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 $q$.

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

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

Using the reference simulation, we can also find:

\[
\Var[J_T] \approx \Var[J_{T, hp}] = 1.1692 \times 10^{-4}
\]

\[
\sigma_g^2 \approx \hat{\sigma}_g^2 = 74.34804 \pm 0.00018,
\]

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

\[
T_d \approx 1.0170 \times 10^{-2}
\]

\[
\sigma_g \approx 8.6225.
\]

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 Figure 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 sampling error, $T_s = N_s \Delta t$ can be increased, holding $\Delta t$ fixed. In Figure 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 $N_s$ towards 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 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 Figure 7 demonstrates is that, in fact, there is a particular optimal scheme in which $\Delta t$ and $T_s$ are simultaneously varied that will extract the most accurate estimate of $J_\infty$ as $N_s$ increases.
FIG. 6: Expected non-dimensional error as a function of non-dimensional timestep for discretizations of the Lorenz equations. $r = 1/2$ assumed.
FIG. 7: Refinement study comparison for fixed $\Delta t$, fixed $T_s$, and optimized $\Delta t$ & $T_s$ using RK3 discretization to compute the expectation of the Lorenz system output $g = u_2$. 
V. INVESTIGATION OF 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:

$$e_{hp} \approx \frac{1}{N_s} \sum_{n=0}^{N_s} \sum_{\eta=n}^{n+Ts/\Delta t} \mathcal{G}(t_\eta, t_n) \circ e_{\text{LT,p}}^{(n)},$$

(31)

where

$$e_{\text{LT,p}}^{(n)} \equiv u_{hp}(t_{n+1}) - u_{\ast}(t_{n+1})$$

(32)

and $u_{\ast}(t_{n+1})$ is exact solution integrated from $u_{hp}(t_n)$ through $\Delta t$:

$$u_{\ast}(t_{n+1}) = u_{hp}(t_n) + \int_{t_n}^{t_{n+1}} f(u_{\ast}(t)) \, dt.$$  

(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:

$$e_{hp} \approx \frac{1}{N_s} \sum_{n=0}^{N_s} \sum_{\eta=n}^{n+Ts/\Delta t} \mathcal{G}(t_\eta, t_n) \circ e_{\text{LT,p}}^{(n)}.$$  

(34)

Now, we assume that a constant $G_{\text{max}}$ exists such that:

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

(35)

for all $t_n, t_\eta \in \mathbb{R}$ and $\mathbf{v} \in B(u(t_n)) \subset \mathbb{R}^d$ where $B(u)$ is the set of states possible by perturbation of $u$ 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 $e_{hp}$:

$$|e_{hp}| \leq \frac{T_d}{\Delta t} G_{\text{max}} \frac{1}{N_s} \sum_{n=0}^{N_s} \left\| e_{\text{LT,p}}^{(n)} \right\|_\infty$$

$$\leq \frac{T_d}{\Delta t} G_{\text{max}} \max_n \left\| e_{\text{LT,p}}^{(n)} \right\|_\infty.$$  

(36)

We now attempt to bound the value of $G_{\text{max}}$ for the Lorenz system by approximating the local truncation error. To make an estimate, we compute both the solution at the
next timestep as well as a surrogate for the true solution at each timestep: \(u_{hp}(t_{n+1})\) and \(\tilde{u}_*(t_{n+1})\), where the former is computed with one timestep 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 \(u_{hp}(t_{n+1})\) and \(\tilde{u}_*(t_{n+1})\) are always advanced from \(u_{hp}(t_n)\). This allows us to estimate \(e^{(n+1)}_{LT,p}\) locally:

\[
e^{(n+1)}_{LT,p} \approx \tilde{e}^{(n+1)}_{LT,p} = u_{hp}(t_{n+1}) - \tilde{u}_*(t_{n+1}).
\]

(37)

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

![Figure 8: Convergence of estimated local truncation error with respect to \(\Delta t\). Fits to \(c_p\Delta t^{p+1}\) shown (with offset for presentation).](image)

Using (36) we can estimate a bounding value for \(G_{\text{max}}\) by

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

(38)

where \(c_p\) is the leading truncation error coefficient fit in Figure 8 and \(C_q\) and \(q\) are taken from Table IIc. Of course when \(q > q_{\text{theory}} = p\), there will be \(\Delta t\) dependence. However, as
\[ \text{(38)} \] requires that the discretization error has an asymptotic behavior, we will only consider $\Delta t$ in the asymptotic convergence regions given in Table I to compute $G_{\text{max}}$. In Figure 9 we show the values of the right-hand side quantity in \( \text{(38)} \), which allow us to make an estimate:

\[ G_{\text{max}} \approx 4.3. \] \( \text{(39)} \)

Next, we use classical truncation error estimates\(^\text{19}\) to relate the discretization error to properties of the solution. We will assume that the local truncation error is bounded by a form:

\[ \max_n \|e_{\text{LT},p}^{(n)}\|_\infty \leq \frac{C_{\text{LT}}}{(p+1)!} \left\| \frac{d^{p+1}u}{dt^{p+1}} \right\|_\infty \Delta t^{p+1} \] \( \text{(40)} \)

where $C_{\text{LT}}$ 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 $u(t)$ can be computed by evaluating $f(u)$ and its derivatives\(^\text{20}\) 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 Figure 10. The resulting values of $C_{\text{LT}}$ that can now be derived by fitting the asymptotic behavior in Figure 8 can be found in Table IV. The result of these estimates is that we can reliably

![FIG. 9: Estimation of bounding value $G_{\text{max}}$.](image-url)
bound the global error of a dynamical system as an accumulation of the local errors over a region of correlation.

![Graph of analytic derivatives](image)

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

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

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

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 with a Hann window function, we perform a spectral analysis on the states of the Lorenz system with a sampling time $T_s = 1000$, $t_0 = 100$, and $\Delta t = 10^{-3}$. The resulting spectrum can be found in Figure 11. We now want to consider how the demonstrated global error behavior demonstrated here results might extrapolate to
more complicated systems by evaluating the spectral behavior of the Lorenz system. Using a discrete Fourier transform with a Hann window function\(^\text{27}\), we perform a spectral analysis on the states of the Lorenz system with a sampling time \(T_s = 1000\), \(t_0 = 100\), and \(\Delta t = 10^{-3}\). The resulting spectrum can be found in Figure [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\(^\text{28}\). 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 Figure [12], which shows the convergence of (28) with the total sampling cost. The effect of increasing order improves the convergence rate in (28) towards the CLT-implied asymptotic rate of \(-1/2\), as well as decreasing the value of the

---

**FIG. 11:** Fourier spectrum of \(u(t)\). Computed with DFT using Hann window function on data from RK4 discretization of Lorenz system with \(T_s = 1000\), \(t_0 = 100\), and \(\Delta t = 10^{-3}\).

Gray dashed line: fit assuming \(|\hat{u}(f)| \approx \exp(-af + b)\) with \(a = 0.872\) and \(b = 2.58\).
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 central limit theorem shown.

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 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; that 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\cite{29}. Consider a Monte Carlo approach to approximate $J_\infty$ with a set of $M_{\text{ens}}$ independent realizations:

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

(41)

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

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

(42)

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

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

(43)

and an optimum given by

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

(44)

at

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

(45)

and

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

(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 Makarashvili et al.\cite{29}; 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 $M_{\text{ens}}$ ensembles with $N_s$ sampling timesteps each will have an equivalent error in expectation to simulating $M_{\text{ens}} N_s$ timesteps in serial. As $M_{\text{ens}}$ is varied on the set of optimal solutions, (45) and (46) indicate that the timestep 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 $u(t)$, a solution of the ergodic chaotic system $f$ from an arbitrary initial condition $u(0) = u_{IC}$ in the basin of attraction of an attractor, $A$. The existence of the attractor implies the non-linear stability of the system, such that all $u_{IC}$ will converge to trajectories on the attractor $A$. Denote by $u^A(t)$ a trajectory that is on the attractor for all $t$ and to which $u(t)$ collapses as $t \to \infty$. The perturbation $\delta u^A(t) \equiv u(t) - u^A(t)$ that describes the IC, therefore, exists in a stable subspace of perturbations to $u^A$ and can be associated with the negative Lyapunov exponents of the system. Thus, we can assume that such perturbations are governed asymptotically by

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

(47)

with $T_\lambda$ a characteristic time associated with the stable Lyapunov modes. In practice, we are interested in averages of quantities on the attractor $g(u^A(t))$, but we can only calculate quantities $g(u(t))$, 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

$$J_T^A = \int_{t_0}^{t_0+T_s} g(u^A(t)) \, dt,$$

(48)

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

$$e_{T, hp} = (J_{T, hp} - J_T) + (J_T - J_T^A) + (J_T^A - J_\infty)$$

(49)

where a new error $e_\lambda$ 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, $A$, 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

24
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 $\epsilon_\lambda$:

$$J_T - J_T^A = \int_{t_0}^{t_0+T_s} (g(u(t)) - g(u^A(t))) \, dt. \tag{50}$$

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

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

will apply for $t \in [0, \infty)$, with $A_\lambda$ a constant that can be related to the deviation between $g(u(0))$ and $g(u^A(0))$.

From this assumption,

$$e_\lambda = \frac{1}{T_s} \int_{t_0}^{t_0+T_s} g(u(t)) - g(u^A(t)) \, dt \approx \frac{1}{T_s} \int_{t_0}^{t_0+T_s} A_\lambda \exp\left(-\frac{t}{T_\lambda}\right) \, dt$$

$$= A_\lambda \frac{T_\lambda}{T_s} \exp\left(-\frac{t_0}{T_\lambda}\right) \left(1 - \exp\left(-\frac{T_s}{T_\lambda}\right)\right). \tag{52}$$

Taking the absolute value, we can find a bounding model:

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

As before, manipulation of (49) allows

$$|e_{T, hp}| = |e_{hp} + e_\lambda + e_T| \leq |e_{hp}| + |e_\lambda| + |e_T|. \tag{54}$$

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

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

where $\mathbb{E}_{B(\mathcal{A})}$ gives the expectation on the basin of attraction of $\mathcal{A}$. Here, the expectation of $|e_{T, hp}|$ doesn’t 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
Because we anticipate $T_\lambda$ will be bounded by a constant for a given system, this is given by:

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

If $A_\lambda$ and $T_\lambda$ can be identified by observation of $g(u(t))$ given an initial condition $u_{IC}$, $|e_\lambda|$ is no longer stochastic and the $\mathbb{E}[e_{T,hp}] \rightarrow |e_{T,hp}|$ 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:

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

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, $e_\lambda$ will scale with the exponent of a large negative value when $t_0 \gg T_\lambda$. Even when $t_0 \gg T_\lambda$, (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 controlling 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 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 $N_{\text{skip}}$ 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 $J_\infty$. The distributions of $g^A(u(t))$ and $g(u(t))$, 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. draws from a normal distribution with mean value $J_\infty$. Then, we have:

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

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 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 $J_\infty$, $\sigma_g$, $A_\lambda$, and $T_\lambda$ in (59) as unknowns, we use Hamiltonian Monte Carlo with the likelihood function implied by (59). We discard from $t = 0$ to $t = 5$, 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 $\{g_n\}$, then use:

\[
\begin{align*}
J_\infty & \sim \mathcal{N}(\tilde{J}, \tilde{\sigma}^2) \\
\sigma_g & \sim \Gamma(\alpha_\sigma, \beta_\sigma) \\
A_\lambda & \sim \mathcal{N}(0, \max(g_{hp}) - \min(g_{hp})) \\
T_\lambda & \sim \Gamma(\alpha_T, \beta_T)
\end{align*}
\] (60)

where

\[
(\alpha_\sigma, \beta_\sigma) \leftarrow \left( \mu_\sigma = \tilde{\sigma}, \sigma_\sigma = \frac{\tilde{\sigma}}{10} \right)
\]

\[
(\alpha_T, \beta_T) \leftarrow (\mu_T = 10.0, \sigma_T = 10.0).
\]

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 be reasonable.

A sample fit and trace are found in Figure 13 for which the maximum a posteriori 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 Figure 13 we can find maximum a posteriori (MAP) estimates of the variables $T_\lambda$ and $|A_\lambda|$ in the decay model. In Figures 14 and 15 histograms of these variables are shown, which are needed to determine (53). We can see that the fit procedure identifies values:

\[
\begin{align*}
T_\lambda & < 4.03 \\
|A_\lambda| & < 38.7
\end{align*}
\] (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. 13: $g = u_2(t)$ trace in transient region, with Bayesian method fit

FIG. 14: MAP estimate $T_\lambda$ for Lorenz system transient. Collected over 1000 Lorenz trajectories with $\Delta t = 10^{-2}$, $T_s = 100$, and randomized $u_{IC}$. Outliers truncated, greater than 97% of data in 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 $u_{IC}$. Outliers truncated, greater than 97% of data in pictured range.
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:

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

With $N$ the total number of timesteps used, given by:

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

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:

$$\left( \frac{e_{model}}{\sigma_g} \right)_{MC} = \frac{A}{\sigma_g T_d} \left( N \left( \frac{\Delta t}{T_d} \right)_{MC} - \frac{t_0}{T_d} \right)^{-1} \exp \left( \frac{t_0/T_d}{A/T_d} \right)$$

$$+ \frac{C_q T_d^q}{\sigma_g} \left( \frac{\Delta t}{T_d} \right)_{MC}^q + \sqrt{\frac{2}{\pi}} M_{ens}^{-\frac{1}{2}} \left( N \left( \frac{\Delta t}{T_d} \right)_{MC} - \frac{t_0}{T_d} \right)^{-\frac{1}{2}}. \quad (64)$$

Using this result, we can solve numerically for $(\Delta t)_{MC,opt}$ and $e_{MC,opt}$ 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 $\Delta t$ and $t_0$ vary with a conservative estimate for the transient behavior using the bounding values in (61). In Figure 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 towards the bottom left, $t_0 \to 0$, $T_s \to 0$, and $\Delta t \ll T_d$; thus the transient error and sampling error become dominant. Similar plots for RK3 and RK4 are found in Figures 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 is make a smaller error possible with a larger timestep, 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.
FIG. 16: Dependence of normalized error expectation $e_{\text{model,MC}}/\sigma_g$ on normalized timestep $\Delta t/T_d$ 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, dashed line indicates optimal $t_0$ given $\Delta t$.

| method | $p$ | $e_{\text{model}}$ | $\Delta t$ | $t_0$ | $T_s$ |
|--------|-----|---------------------|-------------|------|------|
| FE     | 1   | 0.0502              | $2.54 \times 10^{-4}$ | 30.2 | 275  |
| RK3    | 3   | 0.0130              | $8.52 \times 10^{-3}$ | 35.5 | 3370 |
| RK4    | 4   | $8.89 \times 10^{-3}$ | 0.0224      | 36.7 | 6670 |

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

In Figure 19, we take another perspective on these results for RK3 by varying $\Delta t$ and plotting the optimal $t_0$, $T_s$, and $e_{\text{model}}$. As $\Delta t$ 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. 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 $\Delta t$ region, $T_s$ scales with $\Delta t$ both as $\Delta t \to 0$ and as $\Delta t \to \infty$.

The bottom plot of Figure 19 shows the variation of error with $\Delta t$. In this plot we can
FIG. 17: Dependence of normalized error expectation $e_{\text{model,MC}}/\sigma_g$ on normalized timestep $\Delta t/T_d$ and normalized spin-up time $t_0/T_d$ with total cost set at $U = 1.2 \times 10^6$ for 3rd-order Runge Kutta. Red star denotes optimum, dashed line indicates optimal $t_0$ given $\Delta t$.

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 $T_s$ 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 $e_{\text{model,MC}}$ with $U$ under the optimal choices and evaluate how well $e_{\text{model,MC}}$ approximates experimental data for $\mathbb{E}[|J_{MC} - J_\infty|]$. In Figure 20, the variation of $e_{\text{model,MC}}$ computed via (64) as a function of $M_{\text{ens}}$ and $U$ is shown. From this figure, we can see that, in the limit of small error,
FIG. 18: Dependence of normalized error expectation $e_{\text{model,MC}}/\sigma_g$ on normalized timestep $\Delta t/T_d$ and normalized spin-up time $t_0/T_d$ with total cost set at $U = 1.2 \times 10^6$ for 4th-order Runge Kutta. Red star denotes optimum, dashed line indicates optimal $t_0$ given $\Delta t$.

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 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 $M_{\text{ens}}$ and $U$, we generate 1000 individual realizations of $J_{\text{MC}}$ at the computed $(\Delta t)_{\text{MC, opt}}$ and $N_{\text{MC, opt}}$ and using the model fit given in Table IIc. In Figures 21, 22, and 23, we show the predictions and the results of Monte Carlo estimates of $\mathbb{E}[|J_{\text{MC}} - J_{\infty}|]$ for our three discretizations. These results validate the model, with significant discrepancies only when the asymptotic assumptions—$\Delta t$ small and $T_s$ large—do not hold, due to budget limitations in the limit of small $U$. 
FIG. 19: Dependence of normalized spin-up time $t_0/T_d$, sampling time $T_s/T_d$, and model error $e_{\text{model}}/\sigma_g$ on normalized timestep $\Delta t/T_d$ with total cost set at $U = 1.2 \times 10^6$ for 3rd-order Runge Kutta.
FIG. 20: Optimal non-dimensional error under model as a function of total cost $U$ for RK3. Theory totem on left-hand side: discrete convergence rate, $1/q$; on right-hand side: $2(q+r)/q$ rate from (44).

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.

FIG. 23: Total cost model and Monte Carlo validation as a function of total cost $U$ for RK4.
VIII. CONCLUSIONS & FORTHCOMING WORK

In this manuscript, we have developed a theoretical framework for the total error incurred by the discrete sampling of mean outputs 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.

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.\cite{13} can allow us to approximate the model in \textit{(20)} at relatively small cost, and then exploit the result to conduct a high-fidelity simulation at (approximately) optimal discretizations. Further, the framework must be extended to handle chaotic PDE systems as opposed to ODE systems. Though many discrete 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 \textit{and} spatial discretization. These will be the primary concerns of our forthcoming work.

ACKNOWLEDGMENTS

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

CONFLICTS OF INTEREST

The authors have no conflicts of interest to report.

DATA AVAILABILITY

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

1. M. J. Lighthill, “The recently recognized failure of predictability in Newtonian dynamics,” Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences 407, 35–50 (1986).

2. J.-P. Eckmann and D. Ruelle, “Ergodic theory of chaos and strange attractors,” in The theory of chaotic attractors (Springer, 1985) pp. 273–312.

3. D. R. Chapman, “Computational aerodynamics development and outlook,” AIAA journal 17, 1293–1313 (1979).

4. 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,” Advances in DNS/LES 1, 4–8 (1997).

5. H. Choi and P. Moin, “Grid-point requirements for large eddy simulation: Chapman’s estimates revisited,” Physics of Fluids 24, 011702 (2012).

6. J. Kim, P. Moin, and R. Moser, “Turbulence statistics in fully developed channel flow at low Reynolds number,” Journal of Fluid Mechanics 177, 133–166 (1987).

7. A. Lozano-Durán and J. Jiménez, “Effect of the computational domain on direct simulations of turbulent channels up to Reτ = 4200,” Physics of Fluids 26, 011702 (2014), https://doi.org/10.1063/1.4862918.

8. J. C. Del Álamo, J. Jiménez, P. Zandonade, and R. D. Moser, “Scaling of the energy spectra of turbulent channels,” Journal of Fluid Mechanics 500, 135–144 (2004).

9. 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).

10. R. L. Thompson, L. E. B. Sampaio, F. A. de Bragança Alves, L. Thais, and G. Mompéan, “A methodology to evaluate statistical errors in DNS data of plane channel flows,” Computers & Fluids 130, 1–7 (2016).

11. S. Russo and P. Luchini, “A fast algorithm for the estimation of statistical error in DNS (or experimental) time averages,” Journal of Computational Physics 347, 328–340 (2017).

12. 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.
In general, we expect $q = p$, but due to cancellation of local errors, $q > p$ occurs in practice for the Lorenz system. In the expected case of $q = p$, we should expect $G_{\text{max}} = C_q/(c_p T_d)$.

Derivatives of $f$ are computed analytically using the chain rule.
28 G. Karniadakis and S. Sherwin, *Spectral/hp-element methods for computational fluid dynamics* (Oxford University Press, 2005).

29 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,” Computer Physics Communications *219*, 236–245 (2017).