Error behavior and optimal discretization of chaotic differential equations

Cory Frontin

B.S., University of Maryland, College Park (2014)
S.M., Massachusetts Institute of Technology (2018)

Submitted to the Department of Aeronautics & Astronautics in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Aeronautics & Astronautics at the Massachusetts Institute of Technology

February 2023

Author: ....

Department of Aeronautics & Astronautics
7 November 2022

Certified by: ....

David Darmofal
Jerome C. Hunsaker Professor of Aeronautics & Astronautics
Thesis Committee Chair

Certified by: ....

Wesley Harris
Charles Stark Draper Professor of Aeronautics & Astronautics
Thesis Committee Member

Certified by: ....

Jaume Peraire
H.N. Slater Professor of Aeronautics & Astronautics
Thesis Committee Member

Certified by: ....

Qiqi Wang
Associate Professor of Aeronautics & Astronautics
Thesis Committee Member

Accepted by: ....

Jonathan P. How
R.C. Maclaurin Professor of Aeronautics & Astronautics
Chair, Graduate Program Committee

Error behavior and optimal discretization of chaotic differential equations

Cory Frontin

Submitted to the Department of Aeronautics & Astronautics
on 7 November 2022, in partial fulfillment of the requirements for
the degree of Doctor of Philosophy in Aeronautics & Astronautics

Abstract

In this thesis, the simulation of chaotic systems is considered. For many chaotic systems, we desire to make estimates of mean values of quantities of interest, and in this case, the effect of chaos is to introduce behavior that naturally lends itself to statistical, rather than deterministic, description. When simulating chaotic systems using discrete versions of governing differential equations, then, chaos introduces statistical errors alongside discretization errors. These statistical errors are generally one of two types: transient spin-up error before the system reaches the attractor (i.e. the stationary distribution of long-run states) and sampling error due to finite-time averaging of trajectories on the attractor.

In this work, we first propose an error model to describe the expected absolute errors on the attractor of a chaotic ordinary differential equation system. This model for the error implies optimal choices of timestep and sampling time to minimize the error in the simulation– including discretization error and sampling error– given some computational budget. Adding a model for the spin-up error, this allows the description of the optimal choice of timestep, sampling, and spin-up times. Next, we develop a small-sample Bayesian approach that allows the estimation of the discretization and the sampling error using only a small number of simulation results with distinct timesteps and sampling times on the attractor. We then extend the approach for spatiotemporally chaotic partial differential equation systems, which introduces error due to spatial discretization in addition to the temporal discretization errors and statistical errors. Finally, we augment the small-sample approach with corrections for non-negligible spin-up transient behavior, then embed the resulting small-sample method in a naïve explore-exploit algorithm. Using this algorithm, we demonstrate that given a fixed total computational budget such an approach can allow chaotic simulations that achieve near-optimal estimates without strong prior knowledge of the behavior of the system. In addition to this near-optimal discretization, the method allows an a posteriori estimate of the simulation error in the final result after the exploitation stage.

Thesis Supervisor: David Darmofal
Title: Jerome C. Hunsaker Professor of Aeronautics & Astronautics

Error behavior and optimal discretization of chaotic differential equations

Cory Frontin

The following people served as readers for this thesis:

Thesis Reader: ....

Dr. Todd A. Oliver
Research Scientist
Oden Institute
University of Texas

Thesis Reader: ....

Prof. Youssef Marzouk
Professor of Aeronautics & Astronautics
Massachusetts Institute of Technology

Acknowledgments

> *We cannot lose our humanity. It is necessary for us not to lose our solidarity. It is necessary for us to think a little bit with our hearts.*
>
> —Luis Inácio Lula da Silva

To finish a course of Ph.D. research is– not unlike being a human in the first place– to stand on the shoulders of giants. Unfortunately, this also means that it would be impossible to appropriately thank everyone to whom I am indebted for making it possible. This thesis certainly would not have happened without a lot of people laying the groundwork, cleaning up after me, and telling me not to quit, from every possible corner– academic, extracurricular, personal, and familial.

I must start on the academic side of things and thank Dave Darmofal for the support, encouragement, and patience; his example as a researcher is invaluable for its integrity, thoroughness, and–most of all– committment to his collaborators as human beings even before their research output. Shortly thereafter come Dr. Steven Allmaras and Dr. Marshall Galbraith, whose support and friendship throughout my time at MIT has been invaluable. Each member of my committee has also provided invaluable insight: Jaume's teaching in 16.930 crucially formed the basis for much of this research, Qiqi's work in chaotic dynamics piqued my interest in the intersection of chaos with discretization, and Wes's simultaneous committment to excellence and justice has been a guiding light.

The academic life would not be possible without the support of a never-ending list of people. An abbreviated version of that list includes Beth Marois, Jean Sofronas, Robin Courchesne-Sato, Ping Lee, Anthony Zolnik, Beata Schuster, and Pam Fradkin[1]. No list like this would be complete without each and every member of MIT's Facilities Department, who are the silent collaborators with every research project on campus, particularly Emiliano, the regular custodian of the ACDL space for the last few years.

While one learns a lot in the curricular activities of a Ph.D. program, it pales in comparison to what you learn from your peers. Conversations with Michael Brennan and Ricardo Baptista were invalluble to this work as the research crept into domains of knowledge that

Particularly, I appreciate Steve for laughing at so many of my jokes when nobody else was brave enough to admit how funny they were.

I also should add that I am similarly indebted to the readers, Youssef for his passionate pedagogy in 16.940– which I took mostly to do baseball studies only to desperately depend on for this work– and Dr. Oliver, whose work was instrumental in formulating the direction for this one.

Of course, this thesis also would not have been possible without financial support, including support through Research Agreements with Saudi Aramco (under technical monitors Dr. Eric Dow and Dr. Savithru Jayasinghe) and The Boeing Company (under technical monitor Dr. Andrew Cary).

Solidarity to the MIT bargaining unit of SEIU 32BJ!

Also, the person who hand-paints the door markings at MIT, which is one of the small things that has been a source of great joy and contemplation during my time here.

[1] Who also has the inglorius distinction of being my frequent co-conspirator in making good trouble.

they were experts in[2]. I'm super thankful for every member of Dave's research group, notably Philip, Savithru, and Arthur; most of all among these Shun and Ben[3]– my cohort– for their friendship and for everything I learned from them. In the greater ACDL, I owe thanks to too many people to mention them all: I will note Cody, whom I followed here from Maryland's Aerospace Department, and Pat Blonigan who led the welcoming committee when I first got here. In my (too many) extracurricular activities and memberships, I have made far too many friends, in the sense that each and every friendship deserved more attention than I could afford to give it: Stewart, Darien, Cadence, Chelsea, Arthur, and Josué from AeroAfro, BGSA, & GSOC-OC; Mike, Ben, and Davi from the Muddy; Parker, Aaron[4], and Charlotte from GA[3]; Jeff, Carter, Loek, Chris Womack, and all of my friends in the MIT Grad Student Union[5]; also Matt Moraguez, Alex, Aileen, Kevin, and Daniel from the Tech Catholic Community; and last but not least, climbing partners like Emmett, Chris Courtin, and Zack.

Nisha Chandramoorthy was instrumental to the work as I often bothered her about chaos, but more importantly to my life in the lab, through countless deranged and delirious conversations somehow more chaotic than the mathematical ones. I can't thank Hugh enough for immediately informing me on my arrival that I am joining the hockey team and for so warmly welcoming me to the empty seat in "DogeCube", later "HyperCube", only to follow it with years of ongoing friendship. Last but hardly least, thanks to Max Opgenoord for setting an impossibly high standard for success, for entrapping me in the MIT Hyperloop project, and for a great friendship centered in– but extending far beyond– engineering.

I learned a lot from the friends I made along the way, but none of it would have been possible without the folks from home. I am very lucky to have had some great friends stick with me from very early days, Matt & Annette, Logan, Tom, Casto[6], along with a great number more. Thanks for sticking with me even when could be a bit of an absentee friend over the last few years. Most of all, thanks to my family, who have been so supportive even despite my long distance from home. I have especially to thank my grandparents who nurtured my curiousity, my sister who taught me to compete, my late father who taught me to tinker, and my mom, who has always been such a wonderful source of support and love.

[2] Both in the running, alongside Shun, for the title "most excessively nice person in the ACDL".

[3] My sincerest apologies, Ben, as I think I promised you naming rights for my firstborn after you helped me get my 16.920/16.930 code to work. That didn't work out, huh? Hopefully this footnote suffices.

Stewart was also a great roommate, alongside Akil Middleton, who was a fantastic Muddy Charles barfly, which led to our becoming roommates.

As well as the entire set of Muddy Charles Pub staff and barflies!

[4] My goaltending godfather

[5] Solidarity forever with the MITGSU, Local 256 of the United Electrical, Radio, and Machine Workers of America.

I think of the post-Mass hangs in particular, where we, among other things, generalized flat-earth theory to justify gravitation.

This list would also be incomplete without a mention of appreciation my friend and musical idol Eli Roberts.

I also must thank Philippe here, who was no less instrumental in the warm welcome to and in my early years in the lab.

I have not forgotten all of the teachers who have influenced me at various points of my life; I can't possibly name them all, but think particularly of Mary Yarrish, Sam Haller, Jim Roper, and Tom Krazczewicz (hopefully this footnote finally gets me in the Holmes Club).

[6] As well, of course, as the other residents of the Snuggly Duckling: Chris and Yum.

Also, my in-laws, who have tolerated my constant busyness.

Finally, thanks to my wonderful wife Judith and to our Cassius, to both of whom this work is dedicated; I could have done none of it without you.

Verso l'alto.

Contents

Introduction

Introduce a little anarchy.
Upset the established order,
and everything becomes chaos.
—the Joker *(The Dark Knight)*

Since the dawn of aviation, the ability to make predictions about the characteristics of flying machines before they were built differentiated successful aviators from their earthbound counterparts. It is not a coincidence that the Wright Brothers emerged successful in flight in 1903 after building the first wind tunnel capable of experiments accurate enough to assist the process of aircraft design in 1901[7].

Where experimental prediction emerged as a differentiator in the first decades of the 20th century, the open of the 21st might be characterized by the growing industrial use of computational fluid dynamics (CFD) simulations. Since the earliest computational simulation capabilities, CFD has become a key component of the aerospace design lifecycle. The integration of CFD simulations have been shown to have demonstrable benefits in terms of reducing wind tunnel, component, and flight testing times[8]. While these advances are impressive, there remain key challenges for computational fluids in the foreseeable future. Though CFD is a key part modern aerospace design, its present integration into the design lifecycle is limited to "well-behaved" flows, for which reliable and efficient CFD is currently possible.

The advancement of numerical methods for problems in aeronautics through the latter half of the 20th century and the first decades of the 21st was fostered by the exponential rises in processing power under Moore's law and improvements in computational algorithms[9]. As

[7] Michael G. Dodson and David S. Miklosovic. An historical and applied aerodynamic study of the Wright Brothers' wind tunnel test program and application to successful manned flight. In *Fluids Engineering Division Summer Meeting*, volume 1, pages 269–278. American Society of Mechanical Engineers, 06 2005. Symposia, Parts A and B

[8] Jeffrey P Slotnick, Abdollah Khodadoust, Juan Alonso, David Darmofal, William Gropp, Elizabeth Lurie, and Dimitri J Mavriplis. CFD Vision 2030 study: a path to revolutionary computational aerosciences. Technical Report CR–2014-218178, NASA, 2014

[9] Chris A Mack. Fifty years of Moore's law. *IEEE Transactions on Semiconductor Manufacturing*, 24(2):202–207, 2011

an example, this can be seen in the development of top-end channel flow DNS, which has moved from computations with millions of gridpoints with turbulent Reynolds numbers in the hundreds[10] to exascale computations with hundreds of billions of gridpoints and turbulent Reynolds numbers in the mid to high thousands[11]. In spite of progress, significant technical needs remain in order to enable simulations of more complex aerodynamic problems. For example, reliable prediction of drag for an aircraft outside of cruise conditions[12] remains elusive for the foreseeable future.

i.1 Prediction of turbulent flows

In order to make reliable computational predictions for aeronautical questions of increased complexity, methods that resolve the turbulent scales in the problem are necessary. While Reynolds-Averaged Navier-Stokes (RANS) methods work well for predicting steady, attached, and fully turbulent flows, they are well known to be less reliable for predictions when flows are unsteady or detached or transitional[13]. In order to accurately capture these types of flows, we will need to rely on direct numerical simulation (DNS), large eddy simulation (LES) and hybrid RANS-LES approaches.

In a direct numerical simulation, *all* of the relevant physical scales that exist in the problem are resolved. Though it's highly accurate, DNS is expensive beyond usefulness in many engineering-scale applications[14]. In LES, the largest spatial and temporal scales are resolved, while the rest of the scales are modeled using sub-gridscale models[15]. RANS, meanwhile, only resolves an averaged representation of the flow, while modeling all parts of the flow that fluctuate around the average. As a result, the cost of a RANS simulation is signficantly less than an LES.

Original scaling estimates of the cost of modeling a fully turbulent boundary layer with RANS, LES with and without wall modeling, and DNS were made by Chapman[16]. LES costs were also estimated by Spalart, et. al., which improved upon the Chapman results by including the entire boundary layer, from laminar to turbulent, estimating the boundary thickness using RANS[17]. More recent updates of Chapman's estimates have been performed for WMLES, WRLES, and DNS by Choi & Moin[18], although they still did not incorporate the costs of the transitional region. Slotnick, et. al.[19], make estimates for the cost of WMLES, using an integral boundary layer method with a transition model to estimate the boundary layer thickness. The resulting scaling estimates with respect to a

[10] John Kim, Parviz Moin, and Robert Moser. Turbulence statistics in fully developed channel flow at low Reynolds number. *Journal of Fluid Mechanics*, 177:133–166, 1987. DOI: 10.1017/S0022112087000892

[11] Myoungkyu Lee and Robert D. Moser. Direct numerical simulation of turbulent channel flow up to $\mathrm{Re}_\tau \approx$ 5200. *Journal of Fluid Mechanics*, 774: 395–415, 2015

[12] Edward N. Tinoco, Olaf P. Brodersen, Stefan Keye, Kelly R. Laflin, Edward Feltrop, John C. Vassberg, Mori Mani, Ben Rider, Richard A. Wahls, Joseph H. Morrison, David Hue, Christopher J. Roy, Dimitri J. Mavriplis, and Mitsuhiro Murayama. Summary data from the sixth AIAA CFD Drag Prediction Workshop: CRM cases. *Journal of Aircraft*, 55(4):1352–1379, 2018

[13] David Levy, Kelly Laflin, John Vassberg, Edward Tinoco, Mortaza Mani, Ben Rider, Olaf Brodersen, Simone Crippa, Christopher Rumsey, Richard Wahls, Joe Morrison, Dimitri Mavriplis, and Mitsuhiro Murayama. Summary of data from the fifth AIAA CFD Drag Prediction Workshop. In *51st AIAA Aerospace Sciences Meeting*, 2013

[14] Parviz Moin and Krishnan Mahesh. Direct numerical simulations: A tool in turbulence research. *Annual Review of Fluid Mechanics*, 30(1):539–578, 1998

[15] U. Piomelli. Large-eddy simulation: achievements and challenges. *Progress in Aerospace Sciences*, 35(4):335–362, 1999

[16] Dean R. Chapman. Computational aerodynamics development and outlook. *AIAA Journal*, 17(12):1293–1313, 1979

[17] P.R. Spalart, W.H. Jou, M. Strelets, S.R. Allmaras, et al. Comments on the feasibility of LES for wings, and on a hybrid RANS/LES approach. In *Proceedings of the 1st AFOSR Int. Conf. on DNS/LES*, volume 1, pages 4–8, 1997

[18] Haecheon Choi and Parviz Moin. Grid-point requirements for large eddy simulation: Chapman's estimates revisited. *Physics of Fluids*, 24(1):011702, 2012

[19] Slotnick et al., 2014

characteristic Reynolds number, Re, are compiled in Table 1. The key result from these estimates is that the significant costs of LES will very quickly outpace the cost of the RANS approach for complex configurations and these costs are one of the key barriers that must be overcome in order to enable LES to become a more prominent tool in aeronautical design practice.

method	scaling
DNS	$N_x \sim \mathrm{Re}_L^{37/14}$
WRLES	$N_x \sim \mathrm{Re}_L^{13/7}$
WMLES	$N_x \sim \mathrm{Re}_L^{1}$
RANS	$N_x \sim \mathrm{Re}_L^{2/5}$

Table 1: Scaling estimates of gridpoint requirements with respect to characteristic Reynolds number for RANS, LES, and DNS.

Large eddy simulation

In order to understand the primary areas of LES improvement on which we will concentrate, we begin by evaluating the costs associated with LES simulations. An important distinction to begin the study of LES is between wall-resolved LES and wall-modeled LES.

The composition of a turbulent boundary layer can be understood as two main layers. The *inner layer* is the region of the flow nearest the wall, which is relatively predictable and where the most of the turbulent kinetic energy in the flow is produced[20]. The *outer layer*, meanwhile, is much less predictable; while less production of turbulent kinetic energy occurs in the outer layer, it has the effect of convecting the turbulent kinetic energy that is produced by the inner layer.

For wall-resolved LES (WRLES), grid is allocated in order to resolve the large, energy carrying eddies in the outer layer *and* the most significant energy producing eddies of the inner layer. However, the scale of the most important features in the inner layer is significantly smaller than the scale of the most important features in the outer layer[21]. This means that the costs required to perform a WRLES simulation are significant. In wall-modeled LES (WMLES), the relative predictability of the inner layer is exploited by the use of a model for the near wall flow. Thus, in WMLES, the computational effort is concentrated on the outer layer, and the wall model is entrusted with modeling the effect of the inner layer of the flow[22]. As illustrated in Table 1, the cost of resolving near-wall features is much higher than modeling them. However, wall modeling is not yet a mature methodology, and it remains and area of active research[23].

[20] Alexander J. Smits, Beverley J. McKeon, and Ivan Marusic. High–Reynolds number wall turbulence. *Annual Review of Fluid Mechanics*, 43(1):353–375, 2011

[21] Sanjeeb T. Bose and George Ilhwan Park. Wall-modeled large-eddy simulation for complex turbulent flows. *Annual Review of Fluid Mechanics*, 50(1):535–561, 2018

[22] Ugo Piomelli and Elias Balaras. Wall-layer models for large-eddy simulations. *Annual Review of Fluid Mechanics*, 34(1):349–374, 2002

[23] Corentin Carton de Wiart and Scott M. Murman. Assessment of wall-modeled LES strategies within a discontinuous-Galerkin spectral-element framework. In *55th AIAA Aerospace Sciences Meeting*

For wall-modeled LES methods, as well as the related subset of methods known as detached eddy simulation[24] (DES) and other hybrid RANS-LES methods, another crucial area of research is the use of transition models. When transition is not resolved in WMLES and hybrid methods, transition models must be used to dynamically change the modeling choices based on whether or not the model indicates that the flow has transitioned to turbulence[25].

The effect of using a model for small scales in LES and wall models (and, possibly, transition models) in WMLES and hybrid methods will be to introduce *epistemological error* into the simulation. Epistemological error is the form of error that comes from the deficiency of the model to represent reality, and it is a very challenging type of error to quantify. In addition to epistemological error, which can be thought of as the difference between an exact solution to the model equations of motion and the exact true solution, discretization errors exist, which are associated with the effect of numerically estimating solutions to the model equations, and statistical errors, which come from the act of estimating an infinite-time average with a finite one. The proposed work will concentrate on the effects of discretization and statistical error, which we will refer to collectively as "simulation error", and how to minimize them.

i.2 Adaptive discretization methods

A key roadblock for the use of LES and related methods is grid generation. For example, with hybrid methods, it has been shown that accuracy of the method– and its convergence to reference solutions– tends to be frequently strongly dependent on the grids on which the studies are performed[26]. One objective of this thesis to build a framework within which the benefits of mesh adaptivity might be be quantified, understood, and leveraged within the context of turbulent flow simulations. For RANS and other non-chaotic systems, statistical error is not usually a primary concern and the most important factor in the accuracy of a solution to the equations is discretization error. Adaptive methods allow for the optimization of the computational mesh on which the solution is represented in order to control the error in the discretization.

Many modern adaptive finite element methods (FEM) are based on discontinuous Galerkin (DG) and continuous Galerkin (CG) schemes. For clarity, we will concentrate on the DG method here, while noting that extensions of all concepts to continuous discretizations exist[27]. The DG method was developed in order to generalize the family of

[24] Philippe R. Spalart. Detached-eddy simulation. *Annual Review of Fluid Mechanics*, 41(1):181–202, 2009

[25] J. Bodart and J. Larsson. Sensor-based computation of transitional flows using wall-modeled large eddy simulation. In *Annual Research Briefs*, pages 229–240. Center for Turbulence Research, Stanford University, 2012; and George Ilhwan Park and Parviz Moin. An improved dynamic non-equilibrium wall-model for large eddy simulation. *Physics of Fluids*, 26(1):015108, 2014

[26] Spalart, 2009

[27] Hugh A. Carson, Arthur C. Huang, Marshall C. Galbraith, Steven R. Allmaras, and David L. Darmofal. Mesh Optimization via Error Sampling and Synthesis: An update. In *AIAA Scitech Forum*, 2020; and Arthur Chan-wei Huang. *An adaptive variational multiscale method with discontinuous subscales for aerodyanamic flows*. PhD thesis, Massachusetts Institute of Technology, 2020

monotone finite volume (FVM) schemes to formal orders of accuracy greater than two[28]. Thus DG couples the capability of FVM schemes to handle complex geometries with the high-order capability seen in finite difference (FDM) methods. Additionally, DG methods have been seen to be highly parallelizable and particularly well suited for h- and p-adaptation[29]. Ensuing research has also found that DG has very good stability properties, particularly for the solution of Navier-Stokes equations[30]. Furthermore, the DG method can provide high-quality a posteriori error estimates using the dual-weighted residual (DWR) method and the solution to linearized adjoint sensitivity problems[31]. This, in turn, can be exploited to reliably optimize grids via mesh adaptation[32].

Space-time adaptive methods are also possible using a FEM approach. In space-time methods, the spatial *and* temporal dimensions are simultaneously approximated by a finite-element discretization. If the space-time domain is discretized using a $(d+1)$-dimensional discretization, it can then be solved adaptively to minimize the error in the entire spatio-temporal solution[33]. This can allow even more efficient solution than the spatial adaptation technique with temporal time-stepping for problems with concentrated regions of interest in space and time[34].

In order to bring the very promising results for adaptive discretization methods to bear on turbulent simulations, we consider the implications of solving unsteady, turbulent equations of motion, rather than their steady, time-averaged, and non-chaotic counterparts. Foremost among the implications of this transition is how to define "error" when unpredictable and quasi-random behavior is a natural characteristic of the equations of motion. The goal of this thesis is to begin to bridge this gap.

i.3 Chaotic systems

A significant barrier in applying the techniques discussed in the previous section to LES simulations is the fact that solutions from LES and other scale-resolving methods are *chaotic*. Chaotic systems have been characterized[35] as having three primary phenomena:

1. purely deterministic mechanics,
2. aperiodic long-term behavior, and
3. high sensitivity to initial conditions, parametrization, and other perturbations including discretization.

[28] Bernardo Cockburn. *An introduction to the Discontinuous Galerkin method for convection-dominated problems*, pages 150–268. Springer Berlin Heidelberg, Berlin, Heidelberg, 1998. Lectures given at the 2nd Session of the Centro Internazionale Matematico Estivo (C.I.M.E.) held in Cetraro, Italy, June 23–28, 1997

[29] Bernardo Cockburn, George E. Karniadakis, and Chi-Wang Shu. The development of discontinuous Galerkin methods. In Bernardo Cockburn, George E. Karniadakis, and Chi-Wang Shu, editors, *Discontinuous Galerkin Methods*, pages 3–50. Springer Berlin Heidelberg, Berlin, Heidelberg, 2000

[30] Laslo T. Diosady and Scott M. Murman. Higher-order methods for compressible turbulent flows using entropy variables. In *Proceedings of the 53rd AIAA Aerospace Sciences Meeting*, 2015

[31] Roland Becker and Rolf Rannacher. An optimal control approach to a posteriori error estimation in finite element methods. *Acta Numerica*, 10: 1–102, 2001

[32] Masayuki Yano and David L. Darmofal. An optimization-based framework for anisotropic simplex mesh adaptation. *Journal of Computational Physics*, 231(22):7626–7649, 2012; and Masayuki Yano. *An optimization framework for adaptive higher-order discretizations of partial differential equations on anisotropic simplex meshes*. PhD thesis, Massachusetts Institute of Technology, 2012

[33] Thomas J.R. Hughes and Gregory M. Hulbert. Space-time finite element methods for elastodynamics: Formulations and error estimates. *Computer Methods in Applied Mechanics and Engineering*, 66(3):339–363, 1988

[34] Yano, 2012; and Yashod Savithru Jayasinghe. *An adaptive space-time discontinuous Galerkin method for reservoir flows*. PhD thesis, Massachusetts Institute of Technology, 2018

[35] Steven H. Strogatz. *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering*. CRC Press, 2 edition, 2015

Lyapunov stability analysis

A common way to describe the phenomena of chaotic systems mathematically is *Lyapunov stability analysis*. Consider a dynamical system governing $\mathbf{u}(t) : \mathbb{R}^+ \to \mathbb{R}^n$ given by:

$$\frac{\mathrm{d}\mathbf{u}}{\mathrm{d}t} = \mathbf{f}(\mathbf{u}) \qquad \mathbf{u}(0) = \mathbf{u}_0. \tag{i.1}$$

It has been shown that there exist a set of *Lyapunov exponents*

$$\Lambda_1 \geq \Lambda_2 \geq \ldots \geq \Lambda_i \geq \ldots \geq \Lambda_n$$

and *covariant Lyapunov vectors (CLVs)*

$$\psi_1(\mathbf{u}(t)), \psi_2(\mathbf{u}(t)), \ldots, \psi_i(\mathbf{u}(t)), \ldots, \psi_n(\mathbf{u}(t))$$

such that:

$$\frac{\mathrm{d}\psi_i(\mathbf{u}(t))}{\mathrm{d}t} = \left.\frac{\partial \mathbf{f}}{\partial \mathbf{u}}\right|_{\mathbf{u}(t)} \psi_i(\mathbf{u}(t)) - \Lambda_i \psi_i(\mathbf{u}(t)) \tag{i.2}$$

governs perturbations to the dynamics of the system[36]. Using Lyapunov stability theory, it can be shown that, for some $C \in \mathbb{R}^+$:

$$||\delta\mathbf{u}(t)|| \leq C \exp(\Lambda_1 t) \tag{i.3}$$

where $\mathbf{u}(t)$ and $\mathbf{u}(t) + \delta\mathbf{u}(t)$ solve the ODE starting from initial conditions $\mathbf{u}_0$ and $\mathbf{u}_0 + \epsilon_u$, respectively, where ϵ_u is a finite but very small perturbation.

The values of $\{\Lambda_i\}$ and their signs are strongly dependent on the dynamical system $\mathbf{f}$, and they help to describe the long term behavior of the system. If the leading Lyapunov exponent Λ_1 is negative, then the dynamical system will have a single, fixed long-term solution. If Λ_1 is zero, then the system will exhibit a "limit cycle oscillation" with some periodicity. Finally, if $\Lambda_1 > 0$, then the system will be chaotic and have a strange attractor, an infinite set of long-term solutions, assuming that system is bounded. These behaviors have important consequences for both discretization and adaptation, which is of key interest to this work.

Numerical analysis for chaotic systems

Understanding the behavior of discrete solutions of chaotic systems relies upon the numerical analysis for dynamical systems, particularly chaotic ones. With some $t_B > t_A > 0$, the *semigroup operator*[37] $S(\Delta t) : \mathbb{R}^n \to \mathbb{R}^n$ is an operator which evolves the state exactly from t_A to t_B:

$$S(t_B - t_A)\mathbf{u}(t_A) = \mathbf{u}(t_B) = \mathbf{u}(t_A) + \int_{t_A}^{t_B} f(\mathbf{u}(t))\, \mathrm{d}t. \tag{i.4}$$

[36] F. Ginelli, P. Poggi, A. Turchi, H. Chaté, R. Livi, and A. Politi. Characterizing dynamics with Covariant Lyapunov Vectors. *Phys. Rev. Lett.*, 99:130601, Sep 2007

[37] Andrew M. Stuart. Numerical analysis of dynamical systems. *Acta Numerica*, 3:467–572, 1994

For a choice of discretization scheme, we can approximate the evolution of the true solution using a discrete semigroup operator $S_{hp}(\Delta t) : \mathbb{R}^n \to \mathbb{R}^n$:

$$S_{hp}(\Delta t)\mathbf{u}_k = \mathbf{u}_{k+1}. \tag{i.5}$$

which steps the solution from one time-step to another.

In this text, we will denote a discrete variable or operator with a subscript *hp*, which denotes a discretization on a characteristic scale *h* with design order *p*. For ODE systems, *h* translates directly to time-step sizes, i.e. $h_i \iff \Delta t_i$, but we use a more general *h* which will be applicable later to PDE discretizations which have characteristic space *and* time scales: $h_i \iff (\Delta t_i, \Delta x_i)$.

The classical ways of thinking about convergence can be stated using this semigroup notation. For instance, the global L^2 error can be given by:

$$\varepsilon_{hp} = \sqrt{\sum_{k=0}^{N_t} \left[S(k\Delta t)\mathbf{u}^{\mathrm{IC}} - S_{hp}^{k} u_{hp}^{\mathrm{IC}} \right]^2}. \tag{i.6}$$

Based on (i.3), we can expect the inner term to scale with $\exp(\Lambda_1 T)$, where $T = k\Delta t$. For chaotic systems, $\Lambda_1 > 0$. Additionally, if we average over long timescales, T is large. Thus, we can expect the traditional measures of global error to be very large for a chaotic system over long timescales. This matches our phenomenological understanding of chaotic systems: after a relatively short period of time, two solutions that begin from slightly perturbed initial conditions tend to diverge onto completely different trajectories. Unfortunately, these facts mean that we cannot expect global error to be a good measure of usefulness for discrete simulations of chaotic systems.

In order to get at this question, we consider the limiting behavior of solutions in a way that is general enough for chaotic systems, following the work of Stuart[38]. The ω-limit set of a point in the phase space of the solution $\mathbf{u}$ is given by:

$$\omega(\mathbf{u}) = \bigcap_{s \geq 0} \left(\bigcup_{t \geq s} S(t)\mathbf{u} \right), \tag{i.7}$$

associated to a semigroup operator $S(t)$. The ω-limit set, then, describes the set of long-run states that you can get to from $\mathbf{u}$ by the action of $S(t)$. Likewise, the definition generalizes trivially to sets of initial states, $\mathbb{U} = \cup\mathbf{u}$. Using the generalization, we can define the *attractor*, $\mathcal{A}$, which exists when $S(t)\mathbb{U}$ is a uniformly stable attracting set, defined by:

$$\mathcal{A} \equiv \omega(\mathbb{U}) \subseteq \mathbb{U}. \tag{i.8}$$

A consistent discretization and its implied discrete semigroup, S_{hp}, will also have its own ω-limit set, ω_{hp}, and, when $S_{hp}\mathbb{U}_{hp}$ is a uniformly stable asymptotically attracting set, an attractor $\mathcal{A}_{hp}$. In fact, if $S(t)$ has an attractor, then there exists some Δt_c for which

[38] Stuart, 1994

S_{hp} has a stable attracting set $\mathbb{U}_{hp}$ and an attractor $\mathcal{A}_{hp}$ for all $\Delta t \in (0, \Delta t_c]$. We can furthermore write[39] that:

$$\operatorname{dist}(\mathcal{A}_{hp}, \mathcal{A}) \to 0, \tag{i.9}$$

as $\Delta t \to 0$, where $\operatorname{dist}(B, A) = \sup_{u \in B} \inf_{v \in A} \|u - v\|$ is the asymmetric Hausdorff semi-distance.

This is a fairly mathematically dense description, but it illustrates that, while we cannot expect any set of discretizations of a chaotic system to converge to a reference solution, we *can* expect *the attractors* associated with the discretizations to converge toward the true attractor. We can leverage this fact in turn to make meaningful descriptions of the discretization error for simulating chaotic systems.

Ergodicity and long-time outputs of interest

A typical goal of any CFD simulation is to make estimates of one or more outputs of interest, averaged in time, which are functions of the state:

$$J_\infty = \lim_{T \to \infty} \frac{1}{T} \int_{t_0}^{t_0+T_s} g(\mathbf{u}(t)) \, \mathrm{d}t, \tag{i.10}$$

where $g(\mathbf{u}(t))$ are instantaneous measurements like total energy, or lift and drag, which depend in some way on the state, but have direct interest for application to design problems. As we have discussed, $\mathbf{u}_{hp}$ from any two discretizations are likely to be divergent. This implies that the instantaneous output of interest, g, is also likely to be divergent. If $g(\mathbf{u}_{hp})$ does not converge, then, we need to ensure that our approximations of J_∞ can be reasonably compared between discretizations.

Ergodic theory gives us the ability to theoretically justify the comparison of values of J for divergent realizations of $\mathbf{u}_{hp}(t)$ and instantaneous outputs $g(\mathbf{u}_{hp}(t))$. *Ergodicity* describes the tendency of distinct trajectories of a system to converge onto a common attractor after a brief initial transient[40]. In other words, ergodicity is the tendency of a class of systems to "forget" their initial conditions. When a system's state or its outputs are ergodic, then, the state or output can be treated as correlated samples from a stationary distribution: the set of all trajectories on the attractor, after initial transients subside.

The Navier-Stokes equations have been proven in two dimensions to be ergodic[41], and are expected to be ergodic in three dimensions as well. At the very least, three dimensional weak solutions, which we will be concentrating on in the following work, to Navier-Stokes

[39] Jack K. Hale and Geneviéve Raugel. Upper semicontinuity of the attractor for a singularly perturbed hyperbolic equation. *Journal of Differential Equations*, 73(2):197–214, 1988; and Peter E. Kloeden and Jens Lorenz. Stable attracting sets in dynamical systems and in their one-step discretizations. *SIAM Journal on Numerical Analysis*, 23: 986–995, 1986

[40] J.-P. Eckmann and D. Ruelle. Ergodic theory of chaos and strange attractors. In *The Theory of Chaotic Attractors*, pages 273–312. Springer, New York, NY, 2004

[41] Benedetta Ferrario. Ergodic results for stochastic Navier-Stokes equation. *Stochastics*, 60(3-4):271–288, 1997; Franco Flandoli and Bohdan Maslowski. Ergodicity of the 2-D Navier-Stokes equation under random perturbations. *Communications in Mathematical Physics*, 172(1):119–141, 1995; Martin Hairer and Jonathan C. Mattingly. Ergodicity of the 2d Navier-Stokes equations with degenerate stochastic forcing. *Annals of Mathematics*, 164(3):993–1032, 2006; and Jonathan C. Mattingly. Ergodicity of 2D Navier-Stokes equations with random forcing and large viscosity. *Communications in Mathematical Physics*, 206(2):273–288, 1999

are known to converge to a global attractor and be ergodic[42]. The relationship between the existence of weak and the existence of strong solutions of Navier-Stokes in three dimensions are closely related to the famous open mathematical problem of the existence and uniqueness of solutions to the strong form Navier-Stokes equations; nonetheless, we presuppose that the strong form solutions to the three-dimensional Navier-Stokes equations exist and are unique and that the weak form solutions that we will find converge to them, though these presuppositions can not yet be proven. Meanwhile, for the Kuramoto-Sivashinsky equation[43] and the Lorenz equations[44], we can similarly expect the systems to be ergodic. Thus, we can take and compare averages of the form:

$$J_T = \frac{1}{T_s} \int_{t_0}^{t_0+T_s} g(\mathbf{u}) \, \mathrm{d}t. \tag{i.11}$$

meaningfully, so long as T_s is sufficiently large to wipe out the effect of the transient and statistical errors, and, furthermore, we can use these averages to study the convergence of discrete approximations of J_T to J_∞. This idea underpins the error models developed in Chapter 1, we use this framework to develop a model for the error in discrete approximations in terms of the convergence of outputs of interest to the system's true output, J_∞.

It is also worth adding here that we assume that the discrete systems faithfully represent the statistics of the true attractor, though recent work has shown that is not necessarily the case (e.g. Chandramoorthy and Wang [2021]) and further developments may precipitate new requirements to guarantee this is the case.

Finite predictability and unbounded sensitivity for chaotic systems

Before proceeding, we want to highlight a key implication of chaos on the functionality of the adaptive discretization methods that have previously been applied successfully to non-chaotic unsteady flows. One of the early realizations about chaotic systems was their finite predictability[45]: the long-term behavior of chaotic systems are extremely hard to predict given any uncertainty in their initial state. This phenomenon is reflected in (i.3).

For the spacetime adaptive methods, mentioned above, to be applied to chaotic systems, they will require *global solutions* of the spatiotemporal domain, or at least temporal subdivisions of the spatio-temporal domain, which we call *time-slabs*. Typically, the solution algorithm for such a global nonlinear problem (e.g. a Newton method) involves a series of global linearizations. Perturbations to the global linearized solution should grow exponentially for chaotic systems. Thus, if $e^{\Lambda_1 T_{\text{slab}}}$ is large, a global linear system can be expected to suffer from ill-conditioning. Similarly, if an adaptive method relies on an adjoint for its error estimate, the adjoint solution should also be expected to be poorly conditioned. This implies that global solution techniques will be limited to time-slabs of finite and

[42] George R. Sell. Global attractors for the three-dimensional Navier-Stokes equations. *Journal of Dynamics and Differential Equations*, 8(1):1–33, 1996

[43] Patrick J. Blonigan and Qiqi Wang. Least squares shadowing sensitivity analysis of a modified Kuramoto–Sivashinsky equation. *Chaos, Solitons & Fractals*, 64:16–25, 2014

[44] Colin Sparrow. *The Lorenz equations*. Applied Mathematical Sciences. Springer, 1982

[45] Michael James Lighthill, John Michael Tutill Thompson, A. K. Sen, A. G. M. Last, D. J. Tritton, Basil John Mason, P. Mathias, and John Hugh Westcott. The recently recognized failure of predictability in Newtonian dynamics. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, 407(1832):35–50, 1986

limited size such that this amplification factor remains numerically manageable.

Ruelle's linear response theorem is the foundation of sensitivity analysis for chaotic systems and can be expected to hold for the outputs of many of the ergodic chaotic systems of interest[46]. This theorem states that derivatives of outputs of interest with respect to any given parameter s,

$$\frac{\partial J}{\partial s} = \frac{\partial}{\partial s}\left(\int_0^T g(\mathbf{u}(t;s))\right), \tag{i.12}$$

are known to exist and be smooth. However, the *computation* of these values using traditional sensitivity methods, including the adjoint method, break down when there is chaotic sensitivity of the state $\mathbf{u}(t;s)$ to the parameters of interest[47]. This is an area of active research, but while there are some promising methods to calculating sensitivities for chaotic systems[48], sensitivity calculations at practical cost are not anticipated to be possible in the near future, and alternative means of making error estimates for adaptation on chaotic flows must be considered. Both of these issues must be overcome in order to enable mesh adaptation for chaotic flows.

It is outside the scope of this thesis to answer these questions, which are of crucial and parallel importance to the application of this work. In this work, we aim to create a framework in which, first, the errors associated with simulations can be understood in the presence of chaotic dynamics, and second, in which near-optimal choices might be made with respect to these errors. This framework offers to allow the efficacy of discretizations used on chaotic systems– and methods for improving them– to be quantified.

[46] David Ruelle. Differentiation of SRB states for hyperbolic flows. *Ergodic Theory and Dynamical Systems*, 28(2): 613–631, 2008

[47] Qiqi Wang, Rui Hu, and Patrick Blonigan. Least Squares Shadowing sensitivity analysis of chaotic limit cycle oscillations. *Journal of Computational Physics*, 267:210–224, 2014; and Daniel J. Lea, Myles R. Allen, and Thomas W.N. Haine. Sensitivity analysis of the climate of a chaotic system. *Tellus A: Dynamic Meteorology and Oceanography*, 52(5):523–532, 2000

[48] Patrick J. Blonigan, Pablo Fernandez, Scott M. Murman, Qiqi Wang, Georgios Rigas, and Luca Magri. Toward a chaotic adjoint for LES. In *Proceedings of the Summer Program*. Center for Turbulence Research, Stanford University, 2016; and Nisha Chandramoorthy and Qiqi Wang. Sensitivity computation of statistically stationary quantities in turbulent flows. In *AIAA Aviation Forum*, 2019

1

Error behavior for discretizations of ergodic ordinary differential equations

There's no sense in being precise
when you don't even know what you're talking about.

—John von Neumann

For chaotic systems, estimation of long-time behavior is challenging because their governing ordinary differential equations (ODEs) have limited predictability[1]. 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[2]. For ergodic chaotic problems, we frequently want to quantify the unique infinite-time average of some instantaneous quantity of interest of the system:

$$J_\infty = \lim_{T \to \infty} \frac{1}{T} \int_0^T g(\mathbf{u}(t)) \, \mathrm{d}t, \tag{1.1}$$

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

$$\frac{\mathrm{d}\mathbf{u}}{\mathrm{d}t} = \mathbf{f}(\mathbf{u}) \tag{1.2}$$

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

Often, the complexity of a chaotic systems of interest is high, and accordingly the cost of an accurate computational estimate of J_∞ becomes formidable[3]. As the cost of computational simulation gets

[1] Lighthill et al., 1986

[2] Eckmann and Ruelle, 2004

[3] Chapman, 1979; Spalart et al., 1997; and Choi and Moin, 2012

larger, 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 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[4]. Recent work has sought to quantify the effect of statistical error more robustly, using turbulent flow theory[5], advanced spatio-temporal statistical post-processing methods[6], statistical windowing techniques[7], or by extending the concept of Richardson extrapolation to chaotic flows using auto-regressive models and Bayesian methods[8]. The latter work is notable for its use to estimate the statistical errors in the DNS of a high-Re turbulent channel flow[9].

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. [2014], 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 timsteps), 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.

[4] Kim et al., 1987; Adrián Lozano-Durán and Javier Jiménez. Effect of the computational domain on direct simulations of turbulent channels up to $\mathrm{Re}_\tau = 4200$. *Physics of Fluids*, 26 (1):011702, 2014; Juan C. Del Álamo, Javier Jiménez, Paulo Zandonade, and Robert D. Moser. Scaling of the energy spectra of turbulent channels. *Journal of Fluid Mechanics*, 500:135–144, 2004; and Konrad A. Goc, Oriol Lehmkuhl, George Ilhwan Park, Sanjeeb T. Bose, and Parviz Moin. Large eddy simulation of aircraft at affordable cost: a milestone in computational fluid dynamics. *Flow*, 1, 2021

[5] Roney L. Thompson, Luiz Eduardo B. Sampaio, Felipe A.V. de Bragança Alves, Laurent Thais, and Gilmar Mompean. A methodology to evaluate statistical errors in DNS data of plane channel flows. *Computers & Fluids*, 130: 1–7, 2016

[6] Serena Russo and Paolo Luchini. A fast algorithm for the estimation of statistical error in DNS (or experimental) time averages. *Journal of Computational Physics*, 347:328–340, 2017

[7] Charles Mockett, Thilo Knacke, and Frank 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*, pages 9–11. European Research Collaboration on Flow Turbulence and Combustion, 2010

[8] Todd A. Oliver, Nicholas Malaya, Rhys Ulerich, and Robert D. Moser. Estimating uncertainties in statistics computed from direct numerical simulation. *Physics of Fluids*, 26(3): 035101, 2014

[9] Lee and Moser, 2015

1.1 *Proposed error model on the attractor*

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

$$J_{T,hp} = \frac{1}{T_s} \mathrm{I}_{t_0}^{t_0+T_s} \left(g_{hp}(\mathbf{u}_{hp}(t)) \right), \tag{1.3}$$

where the notation $\mathrm{I}_a^b(\cdot)$ 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:

We note here the use of h, which will later be used to denote a spatial and temporal discretization, in which case h will represent both Δt *and* Δx; here, though, it reduces to just Δt.

$$\max_{t \in [0, t_0 + T_s]} \left| g_{hp}(\mathbf{u}_{hp}(t)) - g(\mathbf{u}(t)) \right| = \mathcal{O}(h^p) \tag{1.4}$$

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

$$e_{T,hp} = J_{T,hp} - J_\infty. \tag{1.5}$$

By introducing a third value,

$$J_T = \frac{1}{T_s} \int_{t_0}^{t_0+T_s} g(\mathbf{u}(t))\, \mathrm{d}t, \tag{1.6}$$

we can re-write the error using an identity:

$$e_{T,hp} = (J_{T,hp} - J_T) + (J_T - J_\infty) = e_{hp} + e_T. \tag{1.7}$$

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

$$e_{hp} \equiv J_{T,hp} - J_T \tag{1.8}$$

$$e_T \equiv J_T - J_\infty. \tag{1.9}$$

We can take an absolute value of both sides of (1.7), followed by a manipulation using the triangle inequality:

$$\begin{aligned} |e_{T,hp}| &= |e_{hp} + e_T| \\ &\leq |e_{hp}| + |e_T|. \end{aligned} \tag{1.10}$$

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 $\mathbf{f}$, $\mathcal{A}$, as the set of long-term states towards which all trajectories converge independently of initial condition[10]. 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, $\mathcal{A}$:

$$\mathbb{E}_{\mathcal{A}}[\phi] = \frac{1}{|\mathcal{A}|} \int_{\mathbf{u}_0 \in \mathcal{A}} \phi(\mathbf{u}_0) \, \mathrm{d}\mathbf{u}_0. \tag{1.11}$$

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

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

with, for these examples, $\mathbf{u}(t_0) = \mathbf{u}_0 \in \mathcal{A}$. Given these definitions, we can now take the expectation of (1.10), giving

$$\mathbb{E}_{\mathcal{A}}[|e_{T,hp}|] \leq \mathbb{E}_{\mathcal{A}}[|e_{hp}|] + \mathbb{E}_{\mathcal{A}}[|e_T|] \tag{1.12}$$

by linearity.

From here, we propose asymptotic forms for the two right-hand side terms in (1.12). Consider the definition of e_T in (1.9):

$$e_T = \frac{1}{T_s} \int_{t_0}^{t_0+T_s} g(\mathbf{u}(t)) \, \mathrm{d}t - J_\infty. \tag{1.13}$$

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[11], which has been shown for the Lorenz system[12]. Thus we can write e_T as:

$$e_T \sim \mathcal{N}\left(0, \left(\sqrt{\frac{\pi}{2}} A_0 T_s^{-1/2}\right)^2\right), \tag{1.14}$$

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:

$$|e_T| \sim \mathcal{H}\left(\left(\sqrt{\frac{\pi}{2}} A_0 T_s^{-1/2}\right)^2\right), \tag{1.15}$$

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:

$$\mathbb{E}_{\mathcal{A}}[|e_T|] \approx A_0 T_s^{-1/2} \tag{1.16}$$

[10] Stuart, 1994

[11] Manfred Denker. The central limit theorem for dynamical systems. *Dynamical Systems and Ergodic Theory*, 23:33–62, 1989; and Richard C. Bradley. Basic properties of strong mixing conditions. a survey and some open questions. *Probability Surveys*, 2:107 – 144, 2005

[12] V. Araújo, I. Melbourne, and P. Varandas. Rapid mixing for the Lorenz attractor and statistical limit laws for their time-1 maps. *Communications in Mathematical Physics*, 340(3):901–938, 2015

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 classical analysis[13], we might expect that the discretization error should take a form:

$$|e_{hp}| \approx C_p \left(\frac{\exp(\Lambda T_s) - 1}{\Lambda} \right) (\Delta t)^p. \tag{1.17}$$

This analysis is based on bounding the growth of local truncation error at each timestep by the Lipschitz constant, Λ, of the underlying system, with C_p a constant parameter that depends on the choice of method. However, Viswanath showed[14] that, the global error could be modeled by a form:

$$|e_{hp}| \approx E(T_s; p)(\Delta t)^p, \tag{1.18}$$

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 chaotic systems, the expected convergence onto the attracting set suggests a model of the form:

$$\mathbb{E}_{\mathcal{A}}[|e_{hp}|] \approx C_p(\Delta t)^p, \tag{1.19}$$

As our results in Section 1.2 will show, (1.19) is a good description of the expected discretization error that we observe.

Thus, taking (1.12), (1.16), and (1.19) we assume a bound of the form:

$$\mathbb{E}_{\mathcal{A}}[|e_{T,hp}|] \leq e_{\text{model}} = C_q(\Delta t)^q + A_0 T_s^{-r}, \tag{1.20}$$

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, 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.

1.2 *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[15]:

[13] Ernst Hairer and Gerhard Wanner. Number 14 in Springer Series in Computational Mathematics. Springer, Berlin, Heidelberg, second edition, 1993

[14] Divakar Viswanath. Global errors of numerical ODE solvers and Lyapunov's theory of stability. *IMA Journal of Numerical Analysis*, 21(1):387–406, 01 2001

[15] Edward N. Lorenz. Deterministic nonperiodic flow. *Journal of Atmospheric Sciences*, 20(2):130 – 141, 1963

$$\frac{\mathrm{d}\mathbf{u}}{\mathrm{d}t} = f(\mathbf{u};\boldsymbol{\alpha}) = \begin{pmatrix} \alpha_0(u_1 - u_0) \\ u_0(\alpha_1 - u_2) - u_1 \\ u_0 u_1 - \alpha_2 u_2 \end{pmatrix}, \tag{1.21}$$

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 parametrization[16]: $\boldsymbol{\alpha} = [10, 28, 8/3]$, 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, $p = 1$), 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 $\mathcal{O}(\Delta t^p)$ for non-chaotic systems[17].

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:

$$\mathbf{u}_{\text{init}} \sim \begin{pmatrix} \mathcal{N}(1.0, 5.0^2) \\ \mathcal{N}(1.0, 5.0^2) \\ \mathcal{N}(1.0, 5.0^2) \end{pmatrix} \quad . \tag{1.22}$$

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 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 (1.3) 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 (1.22) and simulated using RK4 with $\Delta t = 17.7 \times 10^{-6}$ and $T_s = 6646.9$. The resulting J_{ref} is:

$$J_{\text{ref}} = 23.549916 \pm 0.000074, \tag{1.23}$$

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 (1.22) and spun-up to independent sampling starting points on the

[16] Sparrow, 1982

[17] J. R. Dormand, R. R. Duckers, and P. J. Prince. Global error estimation with Runge-Kutta methods. *IMA Journal of Numerical Analysis*, 4(2):169–184, 04 1984

attractor $\mathbf{u}_0^{(m)}$:

$$\mathbb{E}[|e_{T,hp}|] \approx \frac{1}{M} \sum_{m=1}^{M} \left| J_{T,hp}\left(\mathbf{u}_0^{(m)}\right) - J_{\mathrm{ref}} \right| . \tag{1.24}$$

In Figures 1.1, 1.2, and 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 1.1. The results of the nonlinear least squares fits for $N_s = 10^4$, 10^5, and 10^6, are given in Table 1.2. 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.

method	$\Delta t_{\max}$	$T_{s,\min}$
FE	5.0×10^{-3}	1.0
RK3	5.0×10^{-2}	1.0
RK4	9.0×10^{-2}	1.0

Table 1.1: Fit boundaries for nonlinear least squares fits.

These figures demonstrate that (1.19) has explanatory value, as the errors in the discretization-dominated region collapse independently of T_s. It is also worth noting that Table 1.2 demonstrates higher-than-expected discretization error convergence rates for FE and RK4.

	FE	RK3	RK4
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

(a) $N_s = 10^4$

	FE	RK3	RK4
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

(b) $N_s = 10^5$

	FE	RK3	RK4
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

(c) $N_s = 10^6$

Table 1.2: Values of error model coefficients computed from nonlinear least squares fits to Monte Carlo study data.

In Figure 1.4, 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 $\Delta t \to 0$.

Finally, we attempt to compare the computational costs across the various discretizations. In this case, the number of timesteps N_s is not

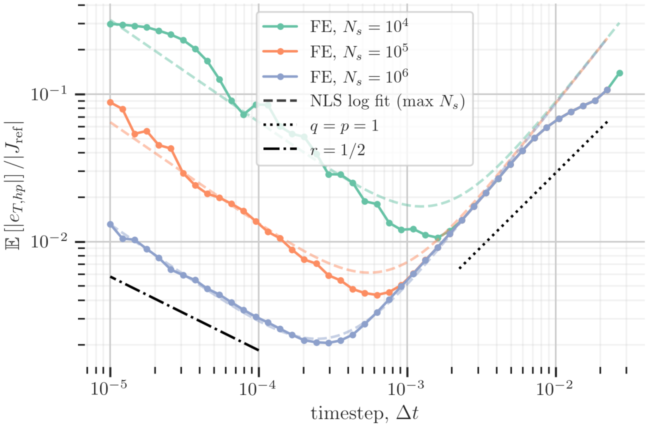


Figure 1.1: Expected relative error as a function of Δt for Forward Euler discretization of the Lorenz equations. Nonlinear least squares fit based on $N_s = 10^6$ data.

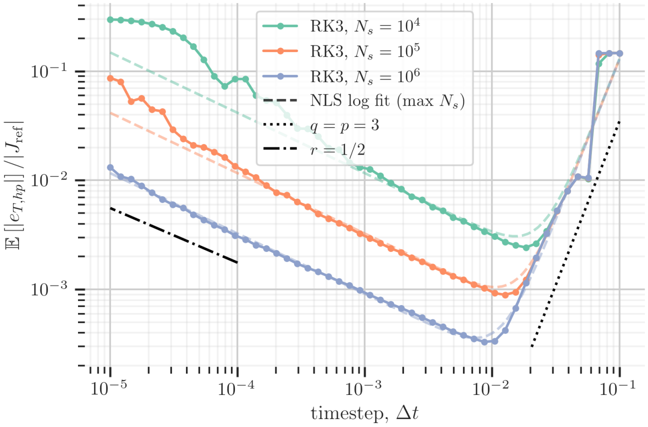


Figure 1.2: Expected relative error as a function of Δt for RK3 discretization of the Lorenz equations. Nonlinear least squares fit based on $N_s = 10^6$ data.

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 $\mathbf{f}$ 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 Figure 1.5, 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 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.

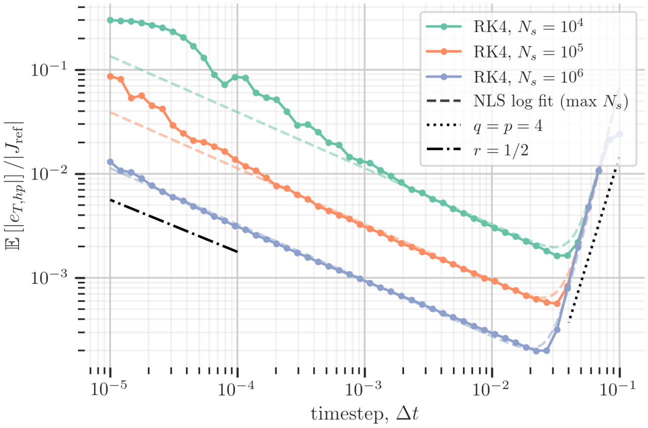


Figure 1.3: Expected relative error as a function of Δt for RK4 discretization of the Lorenz equations. Nonlinear least squares fit based on $N_s = 10^6$ data.

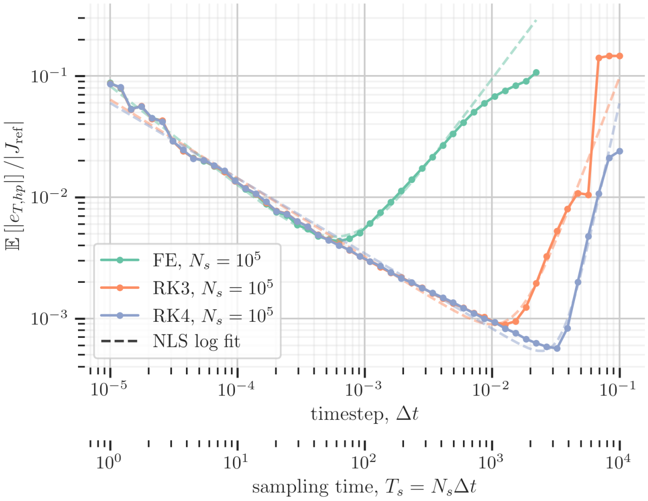


Figure 1.4: Expected relative error as a function of Δt for discretizations of the Lorenz equations.

1.3 *Optimal timestepping on the attractor*

We now study the implications of the error model (1.20), 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 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[18]:

$$\text{Var}[J_T] = \frac{T_d}{T_s}\sigma_g^2. \tag{1.25}$$

[18] Kevin E. Trenberth. Some effects of finite sample size and persistence on meteorological statistics. Part I: Autocorrelations. *Monthly Weather Review*, 112(12):2359 – 2368, 1984

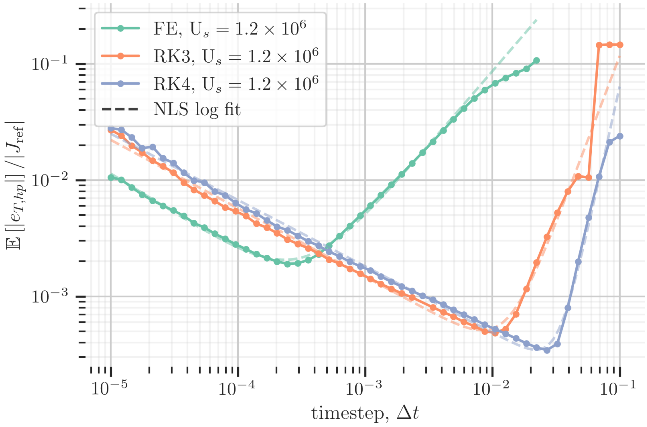


Figure 1.5: Expected percent error as a function of Δt for discretizations of the Lorenz equations at a number of sampling residual evaluations. All fits evaluated at $U_s = 1.2 \times 10^6$.

Furthermore, combining (1.15) and (1.25) allows us to write

$$A_0 = \sqrt{\frac{2}{\pi}} \sigma_g T_d^{1/2}. \tag{1.26}$$

In general, T_d is hard to estimate accurately; this is a crux of the work of Oliver et al. [2014]. 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 Δt and T_s.

The resulting non-dimensional form of the error model is

$$\frac{e_{\mathrm{model}}}{\sigma_g} = \frac{C_q T_d^q}{\sigma_g} \left(\frac{\Delta t}{T_d} \right)^q + \sqrt{\frac{2}{\pi}} \left(\frac{T_s}{T_d} \right)^{-\frac{1}{2}}. \tag{1.27}$$

We can also write the optimizers and optimal value of (1.27) in terms of the non-dimensional variables. These are given by:

$$\begin{aligned} \left(\frac{\Delta t}{T_d} \right)_{\mathrm{opt}} &= \left(\frac{1}{2\pi} \right)^{\frac{1}{2q+1}} \left(\frac{q C_q T_d^q}{\sigma_g} \right)^{-\frac{2}{2q+1}} N_s^{-\frac{1}{2q+1}} \\ \left(\frac{T_s}{T_d} \right)_{\mathrm{opt}} &= \left(\frac{1}{2\pi} \right)^{\frac{1}{2q+1}} \left(\frac{q C_q T_d^q}{\sigma_g} \right)^{-\frac{2}{2q+1}} N_s^{\frac{2q}{2q+1}} \\ \left(\frac{e_{\mathrm{model}}}{\sigma_g} \right)_{\mathrm{opt}} &= \left(\frac{1}{2\pi} \right)^{\frac{q}{2q+1}} \left(2 + \frac{1}{q} \right) \left(\frac{q C_q T_d^q}{\sigma_g} \right)^{\frac{1}{2q+1}} N_s^{-\frac{q}{2q+1}}. \end{aligned} \tag{1.28}$$

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

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

In the limit as $q \to \infty$, the rate $q/(2q+1) \to 1/2$: the CLT limits the convergence rate. Table 1.3 gives the rates of convergence (1.28) for various values of q.

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

Table 1.3: Convergence rates for combined error with respect to sampling timesteps implied by (1.28) at common high-order discretization error convergence rates.

Using the reference simulation, we can also find:

$$\begin{aligned} \mathrm{Var}[J_T] \approx \mathrm{Var}[J_{T,hp}] &= 1.1692 \times 10^{-4} \\ \sigma_g^2 \approx \hat{\sigma}_g^2 &= 74.34804 \pm 0.00018, \end{aligned} \tag{1.29}$$

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

$$\begin{aligned} T_d &\approx 1.0170 \times 10^{-2} \\ \sigma_g &\approx 8.6225. \end{aligned} \tag{1.30}$$

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 Figure 1.6.

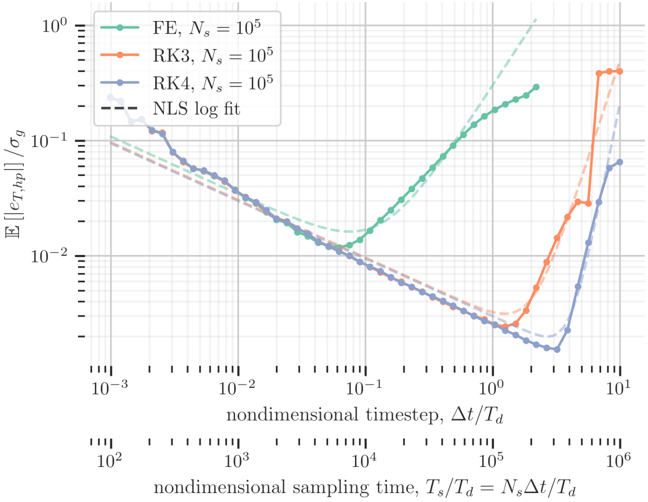


Figure 1.6: Expected non-dimensional error as a function of non-dimensional timestep for discretizations of the Lorenz equations. $r = 1/2$ assumed.

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 Δt fixed. In Figure 1.7, 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: the expected statistical errors prevent more precise 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

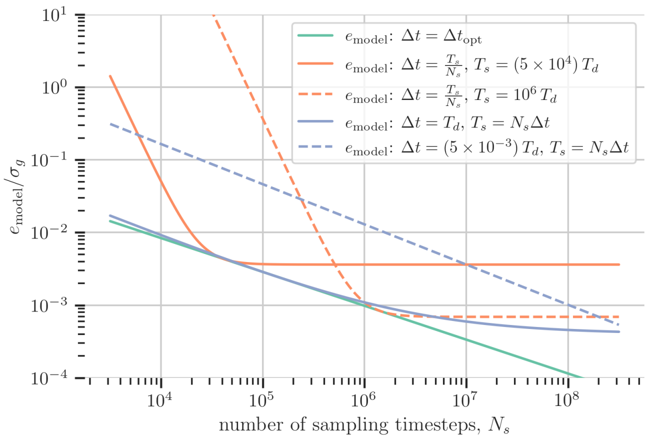


Figure 1.7: Refinement study comparison for fixed Δt, fixed T_s, and optimized Δt & T_s using RK3 discretization to compute the expectation of the Lorenz system output $g = u_2$.

of $1/2$ is initially achieved until the error asymptotes to a constant: discretization errors prevent the more precise 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 (1.20) implies and Figure 1.7 demonstrates is that, in fact, there is a 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.

1.4 *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:

$$e_{hp} \approx \frac{1}{N_s} \sum_{n=0}^{N_s} \sum_{\eta=n}^{N_s} \mathcal{G}(t_\eta, t_n) \circ \mathbf{e}_{\text{LT},p}^{(n)}, \tag{1.31}$$

where

$$\mathbf{e}_{\text{LT,p}}^{(n)} \equiv \mathbf{u}_{hp}(t_{n+1}) - \mathbf{u}_\star(t_{n+1}) \tag{1.32}$$

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

$$\mathbf{u}_\star(t_{n+1}) = \mathbf{u}_{hp}(t_n) + \int_{t_n}^{t_{n+1}} f(\mathbf{u}_\star(t))\, \mathrm{d}t. \tag{1.33}$$

In (1.31), 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:

$$e_{hp} \approx \frac{1}{N_s} \sum_{n=0}^{N_s} \sum_{\eta=n}^{n+T_d/\Delta t} \mathcal{G}(t_\eta, t_n) \circ \mathbf{e}_{\text{LT},p}^{(n)}. \tag{1.34}$$

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

$$\left|\mathcal{G}(t_\eta, t_n) \circ \mathbf{v}\right| \leq \mathcal{G}_{\max} \|\mathbf{v}\|_\infty, \tag{1.35}$$

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

$$\begin{aligned} |e_{hp}| &\leq \frac{T_d}{\Delta t} \mathcal{G}_{\max} \frac{1}{N_s} \sum_{n=0}^{N_s} \left\| \mathbf{e}_{\text{LT},p}^{(n)} \right\|_\infty \\ &\leq \frac{T_d}{\Delta t} \mathcal{G}_{\max} \max_n \left\| \mathbf{e}_{\text{LT},p}^{(n)} \right\|_\infty \end{aligned} \tag{1.36}$$

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 timestep as well as a surrogate for the true solution at each timestep: $\mathbf{u}_{hp}(t_{n+1})$ and $\tilde{\mathbf{u}}_\star(t_{n+1})$, 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 $\mathbf{u}_{hp}(t_{n+1})$ and $\tilde{\mathbf{u}}_\star(t_{n+1})$ are always advanced from $\mathbf{u}_{hp}(t_n)$. This allows us to estimate $\mathbf{e}_{\text{LT},p}^{(n+1)}$ locally:

$$\mathbf{e}_{\text{LT},p}^{(n+1)} \approx \tilde{\mathbf{e}}_{\text{LT},p}^{(n+1)} = \mathbf{u}_{hp}(t_{n+1}) - \tilde{\mathbf{u}}_\star(t_{n+1}). \tag{1.37}$$

In Figure 1.8 we characterize the convergence of local error estimates. Computations are run with $T_s = 100$ and $t_0 = 100$ fixed, varying Δt. At each timestep, the local truncation error is estimated by computing (1.37). The figure shows the computed $\max_n ||\tilde{\mathbf{e}}_{\text{LT},p}^{(n)}||_\infty$ and demonstrates that the expected rate of $(p+1)$ is nearly exactly achieved.

Using (1.36) we can estimate a bounding value for $\mathcal{G}_{\max}$ by

$$\mathcal{G}_{\max} \geq \frac{\mathbb{E}[|e_{hp}|]}{\max_n \left\| \mathbf{e}_{\text{LT},p}^{(n)} \right\|_\infty} \frac{\Delta t}{T_d} = \frac{C_q \Delta t^q}{c_p \Delta t^{p+1}} \frac{\Delta t}{T_d}, \tag{1.38}$$

where c_p is the leading truncation error coefficient fit in Figure 1.8, and C_q and q are taken from Table 1.2. Of course when $q > q_{\text{theory}} = p$,

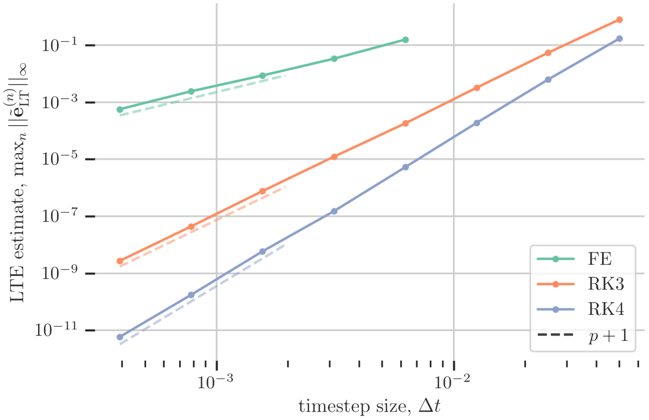


Figure 1.8: Convergence of estimated local truncation error with respect to Δt. Fits to $c_p \Delta t^{p+1}$ shown (with offset for presentation).

there will be Δt dependence. However, as (1.38) requires that the discretization error has an asymptotic behavior, we will only consider Δt in the asymptotic convergence regions given in Table 1.1 to compute $\mathcal{G}_{\max}$. In Figure 1.9, we show the values of the right-hand side quantity in (1.38), which allow us to make an estimate:

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 $\mathcal{G}_{\max} = C_q/(c_p T_d)$.

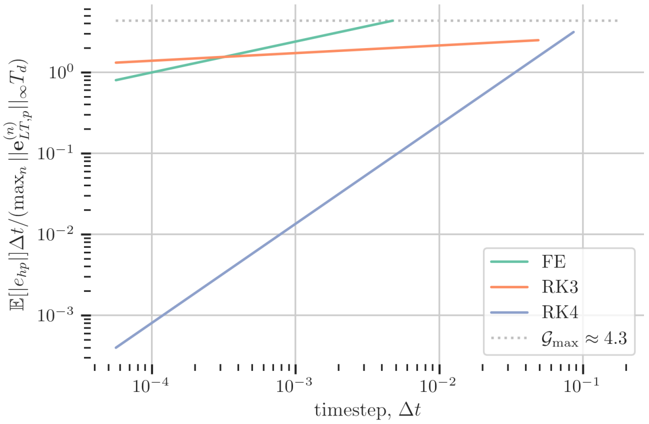


Figure 1.9: Estimation of bounding value $\mathcal{G}_{\max}$.

$$\mathcal{G}_{\max} \approx 4.3. \tag{1.39}$$

Next, we use classical truncation error estimates[19] to relate the discretization error to properties of the solution. We will assume that the local truncation error is bounded by a form:

$$\max_n \left\| \mathbf{e}^{(n)}_{\mathrm{LT},p} \right\|_\infty \leq \frac{C_{\mathrm{LT}}}{(p+1)!} \left\| \frac{\mathrm{d}^{p+1}\mathbf{u}}{\mathrm{d}t^{p+1}} \right\|_\infty \Delta t^{p+1} \tag{1.40}$$

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

[19] Hairer and Wanner, 1993

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 Figure 1.10. The resulting values of C_{LT} that can now be derived by fitting the asymptotic behavior in Figure 1.8 can be found in Table 1.4. The result of these estimates is that we can reliably bound the global error of a dynamical system as an accumulation of the local errors over a region of correlation.

Derivatives of $\mathbf{f}$ are computed analytically using the chain rule.

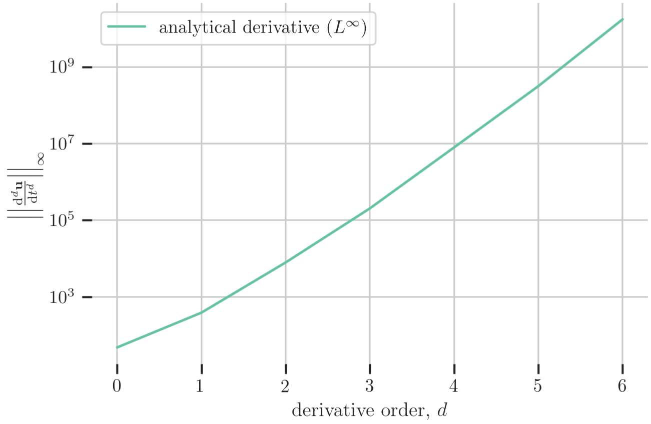


Figure 1.10: Norm of analytic derivatives of $\mathbf{u}$ computed on the attractor of $\mathbf{f}$. State $\mathbf{u}$ computed with RK4 at $\Delta t = 10^{-4}$ and $T_s = 1000$ after discarding $t_0 = 100$.

p	rate (observed)	$C_{\mathrm{LT}}\left\|\frac{\mathrm{d}^{p+1}\mathbf{u}}{\mathrm{d}t^{p+1}}\right\|_\infty$	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

Table 1.4: Rate and coefficient fit for convergence of local truncation error of discrete Lorenz system. $C_{\mathrm{LT}}\left\|\frac{\mathrm{d}^{p+1}\mathbf{u}}{\mathrm{d}t^{p+1}}\right\|_\infty$ estimated by $c_p(p+1)!$ using c_p fit from Figure 1.8.

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[20], 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 1.11. The Lorenz system tends to have the most content in the frequencies with $f \lesssim 10^1$, with a region

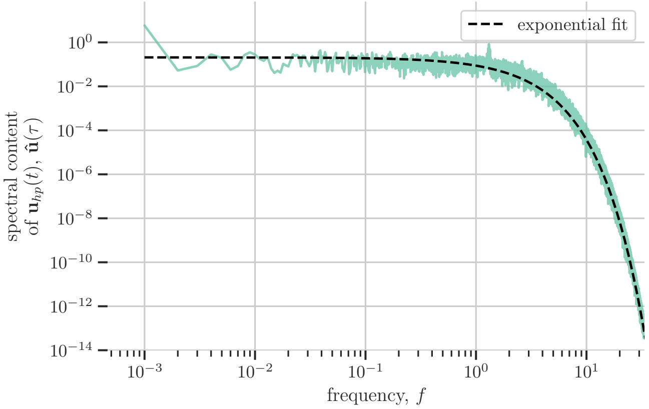


Figure 1.11: Fourier spectrum of $\mathbf{u}(t)$. Computed with DFT using Hann window function on data from RK4 discretization of Lorenz system with $T_s = 1000$, $t_0 = 100$, and $\Delta t = 10^{-3}$. Gray dashed line: fit assuming $|\hat{\mathbf{u}}(f)| \approx \exp(-af + b)$ with $a = 0.872$ and $b = 2.58$.

[20] F.J. Harris. On the use of windows for harmonic analysis with the discrete Fourier transform. *Proceedings of the IEEE*, 66(1):51–83, 1978

of exponential decay in the range $1 \lesssim f \lesssim 30$. On scales with $f \gtrsim 30$, machine precision plateaus are observed and omitted here.

The fact that the Lorenz spectrum is an exponentially decreasing function of frequency $\mathbf{f}$ makes the use of high-order methods theoretically appealing for the spectral convergence of hp-refinement strategies[21]. Unfortunately, the effect of statistical error in (1.28) 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 1.12, which shows the convergence of (1.28) with the total sampling cost. The effect of increasing order improves the convergence rate in (1.28) towards the CLT-implied asymptotic rate of $-1/2$, as well as decreasing the value of the leading constant, but 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}.

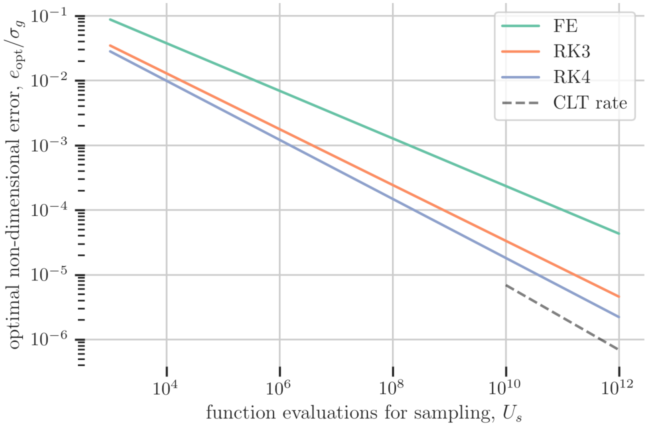


Figure 1.12: Convergence of optimal error with sampling costs for FE, RK3, and RK4 discretizations of the Lorenz output $g = u_2$. Asymptotic $1/2$ rate implied by central limit theorem shown.

1.5 *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.

[21] George Karniadakis and Spencer Sherwin. *Spectral/hp element methods for computational fluid dynamics*. Oxford University Press, 06 2005

Ensemble averaging on the attractor

It is well understood how sampling error can be reduced at a fixed wall clock time by ensemble averaging across multiple parallel processes[22]; we now consider the effect of ensemble averaging when the effect of discretization error is included. Consider a Monte Carlo approach to approximate J_∞ with a set of M_ens independent realizations:

$$J_\text{MC} = \frac{1}{M_\text{ens}} \sum_{m=1}^{M_\text{ens}} J_{T,hp}^{(m)}. \tag{1.41}$$

We can write a modified version of (1.20) to approximate the error that we expect in the Monte Carlo estimator in (1.41):

$$\mathbb{E}[|J_\text{MC} - J_\infty|] \approx e_\text{model,MC} = C_q (\Delta t)_\text{MC}^q + \frac{A_0}{\sqrt{M_\text{ens}}} T_{s,\text{MC}}^{-r}, \tag{1.42}$$

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

$$\left(\frac{e_\text{model}}{\sigma_g}\right)_\text{MC} = \frac{C_q T_d^q}{\sigma_g} \left(\frac{\Delta t}{T_d}\right)^q + \sqrt{\frac{2}{\pi}} M_\text{ens}^{-\frac{1}{2}} \left(\frac{T_s}{T_d}\right)^{-\frac{1}{2}}, \tag{1.43}$$

and an optimum given by

$$\left(\frac{e_\text{model}}{\sigma_g}\right)_\text{MC,opt} = \left(\frac{1}{2\pi}\right)^{\frac{q}{2q+1}} \left(2 + \frac{1}{q}\right) \left(\frac{q C_q T_d^q}{\sigma_g}\right)^{\frac{1}{2q+1}} M_\text{ens}^{-\frac{q}{2q+1}} N_s^{-\frac{q}{2q+1}}, \tag{1.44}$$

at

$$\left(\frac{\Delta t}{T_d}\right)_\text{opt} = \left(\frac{1}{2\pi}\right)^{\frac{1}{2q+1}} \left(\frac{q C_q T_d^q}{\sigma_g}\right)^{-\frac{2}{2q+1}} M_\text{ens}^{-\frac{1}{2q+1}} N_s^{-\frac{1}{2q+1}}, \tag{1.45}$$

and

$$\left(\frac{T_s}{T_d}\right)_\text{opt} = \left(\frac{1}{2\pi}\right)^{\frac{1}{2q+1}} \left(\frac{q C_q T_d^q}{\sigma_g}\right)^{-\frac{2}{2q+1}} M_\text{ens}^{-\frac{1}{2q+1}} N_s^{\frac{2q}{2q+1}}. \tag{1.46}$$

Equation 1.44 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. [2017]; 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 simulating $M_\text{ens} N_s$ timesteps in serial. As M_ens is varied on the set of optimal solutions, (1.45) and (1.46) indicate that the timestep and sampling time should be reduced with the same factor $M_\text{ens}^{-1/(2q+1)}$ as M_ens increases in order to achieve perfect scaling.

[22] Vakhtang Makarashvili, Elia Merzari, Aleksandr Obabko, Andrew Siegel, and Paul Fischer. A performance analysis of ensemble averaging for high fidelity turbulence simulations at the strong scaling limit. *Computer Physics Communications*, 219:236–245, 2017

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 $\mathbf{f}$ from an arbitrary initial condition $\mathbf{u}(0) = \mathbf{u}_{\mathrm{IC}}$ in the basin of attraction of an attractor, $\mathcal{A}$. The existence of the attractor implies the non-linear stability of the system, such that all $\mathbf{u}_{\mathrm{IC}}$ 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 $t \rightarrow \infty$. 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

$$\left\| \delta\mathbf{u}^{\mathcal{A}}(t) \right\| \lesssim \exp\left(-\frac{t}{T_\lambda} \right), \tag{1.47}$$

with T_λ 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 our earlier idenitication of (1.7), we have effectively found an estimate of

$$J_T^{\mathcal{A}} = \frac{1}{T_s} \int_{t_0}^{t_0+T_s} g(\mathbf{u}^{\mathcal{A}}(t))\, \mathrm{d}t, \tag{1.48}$$

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

$$e_{T,hp} = \underbrace{(J_{T,hp} - J_T)}_{e_{hp}} + \underbrace{(J_T - J_T^{\mathcal{A}})}_{e_\lambda} + \underbrace{(J_T^{\mathcal{A}} - J_\infty)}_{e_T} \tag{1.49}$$

where a new error e_λ is introduced, associated with the spin-up transient. The model for e_T in (1.9) will apply without modification, while the model for e_{hp} will be subject to slightly different assumptions. Where in (1.8), 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

We note here that there is some arbitrariness in the ordering of this equation. Both the true system and the discretized ones will have a tendency to have transient behavior before arriving at a stationary distribution. We can account for the transient behavior (as we have here) as being a characteristic of the finite-time true system, or as a characteristic of the discrete, finite time system, such that e_{hp} is just on the attractor. In terms of the equations, this would take the form: $e_\lambda \equiv J_{T,hp} - J_{T,hp}^{\mathcal{A}}$, $e_{hp} \equiv J_{T,hp}^{\mathcal{A}} - J_T^{\mathcal{A}}$, and e_T as before. This latter description is likely to be rather useful: it's more theoretically succinct and represents a more readily estimable quantity; however, we use the former in this work since we estimate the expected transient behavior using reference simulations, which approximates the behavior of the true system.

$\mathcal{A}$, $B(\mathcal{A})$. We assume that a model of the form used in (1.8) applies in expectation when the transient component is included.

Next, we concentrate on e_λ:

$$J_T - J_T^{\mathcal{A}} = \int_{t_0}^{t_0+T_s} \Big(g(\mathbf{u}(t)) - g(\mathbf{u}^{\mathcal{A}}(t)) \Big) \ \mathrm{d}t. \tag{1.50}$$

We now assume that, like $\mathbf{u}$, g will decay exponentially in t as (1.47), such that

$$g(\mathbf{u}(t)) - g(\mathbf{u}^{\mathcal{A}}(t)) \equiv \delta g^{\mathcal{A}}(t) \approx A_\lambda \exp\left(-\frac{t}{T_\lambda} \right) \tag{1.51}$$

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

From this assumption,

$$\begin{aligned} e_\lambda &= \frac{1}{T_s} \int_{t_0}^{t_0+T_s} g(\mathbf{u}(t)) - g(\mathbf{u}^{\mathcal{A}}(t)) \ \mathrm{d}t \\ &\approx \frac{1}{T_s} \int_{t_0}^{t_0+T_s} A_\lambda \exp\left(-\frac{t}{T_\lambda} \right) \ \mathrm{d}t \\ &= A_\lambda \frac{T_\lambda}{T_s} \exp\left(-\frac{t_0}{T_\lambda} \right) \left(1 - \exp\left(-\frac{T_s}{T_\lambda} \right) \right) \\ &\approx A_\lambda \frac{T_\lambda}{T_s} \exp\left(-\frac{t_0}{T_\lambda} \right) \equiv e_\lambda \end{aligned} \tag{1.52}$$

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

$$|e_\lambda| = |A_\lambda| \frac{T_\lambda}{T_s} \exp\left(-\frac{t_0}{T_\lambda} \right). \tag{1.53}$$

As before, manipulation of (1.49) allows

$$|e_{T,hp}| = |e_{hp} + e_\lambda + e_T| \tag{1.54}$$

$$\leq |e_{hp}| + |e_\lambda| + |e_T|. \tag{1.55}$$

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

$$\mathbb{E}[|e_{T,hp}|] \leq \mathbb{E}_{B(\mathcal{A})}[|e_{hp}|] + \mathbb{E}_{\mathrm{IC}}[|e_\lambda|] + \mathbb{E}_{\mathcal{A}}[|e_T|], \tag{1.56}$$

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 (1.53) to complete (1.56). Because we anticipate a constant T_λ will be bounded for a given system, this is given by:

$$\mathbb{E}_{\mathrm{IC}}[|e_\lambda|] = \mathbb{E}_{\mathrm{IC}}[|A_\lambda|]\frac{T_\lambda}{T_s}\exp\left(-\frac{t_0}{T_\lambda}\right). \tag{1.57}$$

If a A_λ and T_λ can be identified by observation of $g(\mathbf{u}(t))$ given an initial condition $\mathbf{u}_{\mathrm{IC}}$, $|e_\lambda|$ is no longer stochastic and the $\mathbb{E}[|e_{T,hp}|] \to |e_{T,hp}|$ as in (1.53).

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

$$e_{\mathrm{model,MC}} = \tilde{A}_\lambda \frac{T_\lambda}{T_{s,\mathrm{MC}}}\exp\left(-\frac{t_0}{T_\lambda}\right) + C_q(\Delta t)^q_{\mathrm{MC}} + \frac{A_0}{\sqrt{M_{\mathrm{ens}}}}T^{-r}_{s,\mathrm{MC}}, \tag{1.58}$$

where $\tilde{A}_\lambda$ can be either estimated on an instance-by-instance basis or by estimating the expectation on a 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 $t_0 \not\gg T_\lambda$, (1.53) 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 as $T_s \gg 1$. This also implies two "paths" to controlling spin-up errors: either choosing t_0 long enough to shrink the mean offset error to negligibility at $t = t_0$, or choosing T_s long enough so that the mean offset contribution to the simulation error is small in spite of the error at $t = t_0$.

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 $g_n^{\mathcal{A}} \equiv g^{\mathcal{A}}(t_n)$ for t_n in $\{t_0, t_0 + N_{\mathrm{skip}}\Delta t, \ldots, t_0 + T_s\}$. 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 $g_n^{\mathcal{A}}$ will be an independent and identically distributed (i.i.d.) draw from a bounded, stationary distribution with mean J_∞. The distributions of $g^{\mathcal{A}}(\mathbf{u}(t))$ and $g(\mathbf{u}(t))$, in general, are not known. In order to facilitate an estimate of the mean behavior, we will assume $g_n^{\mathcal{A}}$ are i.i.d. draws from a normal distribution with mean value J_∞. Then, we have:

$$g_n \sim \mathcal{N}(J_\infty + \delta g^{\mathcal{A}}(t_n), \sigma_g^2), \tag{1.59}$$

where the relationship between g_n and $g_n^{\mathcal{A}}$ is taken from (1.51).

In order to understand the implications of this model, we can use the 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 (1.22), 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 (1.59) as unknowns, we use Hamiltonian Monte Carlo with the likelihood function implied by (1.59). 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:

$$\begin{aligned} J_\infty &\sim \mathcal{N}(\tilde{J}, \tilde{\sigma}^2) \\ \sigma_g &\sim \Gamma(\alpha_\sigma, \beta_\sigma) \\ A_\lambda &\sim \mathcal{N}(0, \max(g_{hp}) - \min(g_{hp})) \\ T_\lambda &\sim \Gamma(\alpha_T, \beta_T) \end{aligned} \tag{1.60}$$

where

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

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 be reasonable.

A sample fit and trace are found in Figure 1.13, for which the maximum a posteriori estimate gives $T_\lambda = 0.312$ and $A_\lambda = -0.925$.

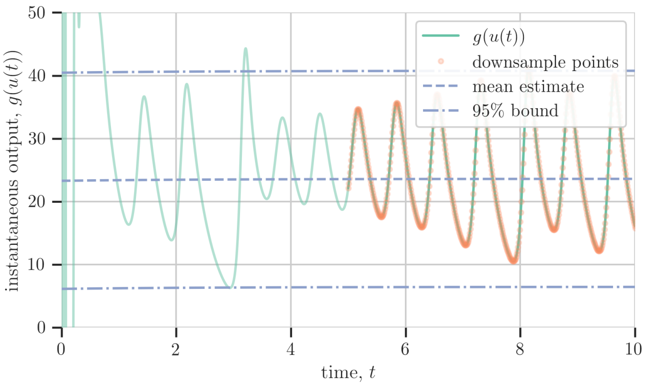


Figure 1.13: $g = u_2(t)$ trace in transient region, with Bayesian method fit

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 1.13 we can find maximum a posteriori (MAP) estimates of the variables T_λ and $|A_\lambda|$ in the decay model. In Figures 1.14 and 1.15, histograms of these variables are shown, which are needed to determine (1.53). We can see that the fit procedure identifies values:

$$\begin{aligned} T_\lambda &< 4.03 \\ |A_\lambda| &< 38.7 \end{aligned} \tag{1.61}$$

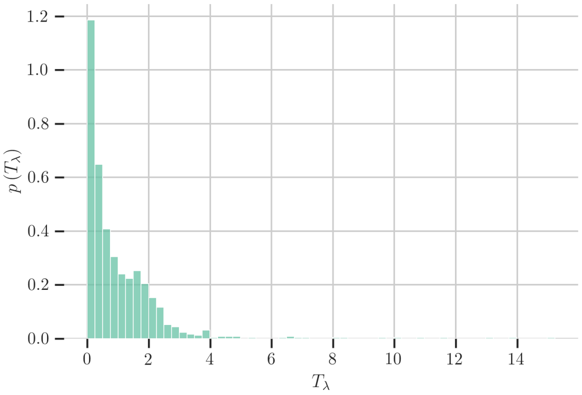


Figure 1.14: MAP estimate T_λ for Lorenz system transient. Collected over 1000 Lorenz trajectories with $\Delta t = 10^{-2}$, $T_s = 100$, and randomized $\mathbf{u}_{\mathrm{IC}}$. Outliers truncated, greater than 97% of data in pictured range.

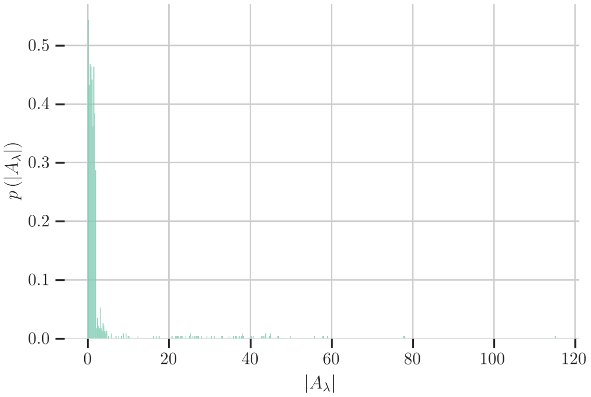


Figure 1.15: MAP estimate $|A_\lambda|$ for Lorenz system transient. Collected over 1000 Lorenz trajectories with $\Delta t = 10^{-2}$, $T_s = 100$, and randomized $\mathbf{u}_{\mathrm{IC}}$. Outliers truncated, greater than 97% of data in pictured range.

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.

1.6 *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:

$$N_0 = \left\lceil \frac{t_0}{(\Delta t)_{\mathrm{MC}}} \right\rceil \approx \frac{t_0}{(\Delta t)_{\mathrm{MC}}}. \tag{1.62}$$

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

$$N = N_0 + N_{\mathrm{MC}} = \frac{t_0}{(\Delta t)_{\mathrm{MC}}} + N_{\mathrm{MC}}, \tag{1.63}$$

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

By normalizing (1.58) then substituting (1.63), we arrive at a transient-inclusive non-dimensional model for the error:

$$\left(\frac{e_{\text{model}}}{\sigma_g}\right)_{\text{MC}} = \frac{\tilde{A}_\lambda}{\sigma_g}\frac{T_\lambda}{T_d}\left(N\left(\frac{\Delta t}{T_d}\right)_{\text{MC}} - \frac{t_0}{T_d}\right)^{-1}\exp\left(-\frac{t_0/T_d}{T_\lambda/T_d}\right) + \frac{C_q T_d^q}{\sigma_g}\left(\frac{\Delta t}{T_d}\right)^q_{\text{MC}} + \sqrt{\frac{2}{\pi}}M_{\text{ens}}^{-\frac{1}{2}}\left(N\left(\frac{\Delta t}{T_d}\right)_{\text{MC}} - \frac{t_0}{T_d}\right)^{-\frac{1}{2}}. \tag{1.64}$$

Using this result, we can solve numerically for $(\Delta t)_{\text{MC,opt}}$ and $e_{\text{MC,opt}}$ via (1.64).

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 (1.64) as Δt and t_0 vary with a conservative estimate for the transient behavior using the bounding values in (1.61). In Figure 1.16, we show e_{model}

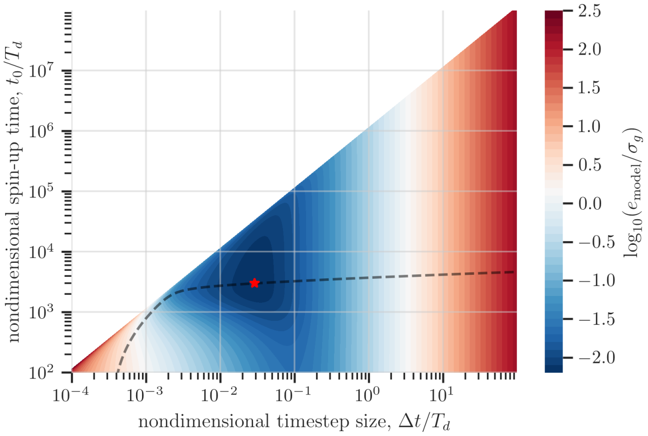


Figure 1.16: Dependence of normalized error expectation $e_{\text{model,MC}}/\sigma_g$ on normalized timestep $\Delta t/T_d$ and normalized spin-up time t_0/T_d with total cost set at $U = 1.2 \times 10^6$ for Forward Euler. Red star denotes optimum, dashed line indicates optimal t_0 given Δt.

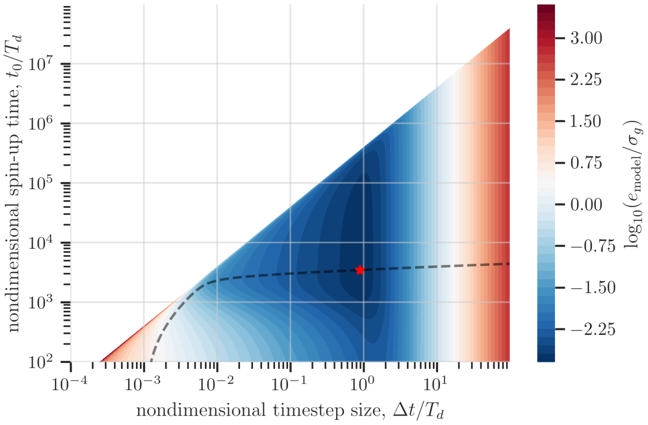


Figure 1.17: Dependence of normalized error expectation $e_{\text{model,MC}}/\sigma_g$ on normalized timestep $\Delta t/T_d$ and normalized spin-up time t_0/T_d with total cost set at $U = 1.2 \times 10^6$ for 3^{rd}-order Runge Kutta. Red star denotes optimum, dashed line indicates optimal t_0 given Δt.

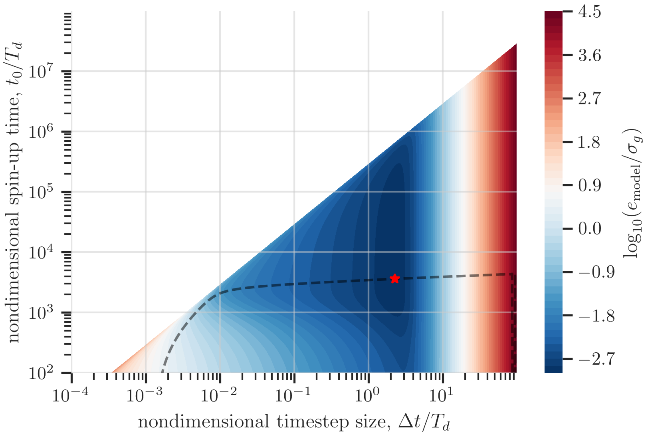


Figure 1.18: Dependence of normalized error expectation $e_{\mathrm{model,MC}}/\sigma_g$ on normalized timestep $\Delta t/T_d$ and normalized spin-up time t_0/T_d with total cost set at $U = 1.2 \times 10^6$ for 4th-order Runge Kutta. Red star denotes optimum, dashed line indicates optimal t_0 given Δt.

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 1.17 and 1.18. The optimal errors and optimizing simulations are described in Table 1.5. 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.

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

Table 1.5: Optimal Lorenz simulations for output $g = u_2$ under budget of $U = 1.2 \times 10^6$ right-hand side evaluations using $M_{\mathrm{ens}} = 1$.

In Figure 1.19, 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. 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 $\Delta t \to 0$ and as $\Delta t \to \infty$.

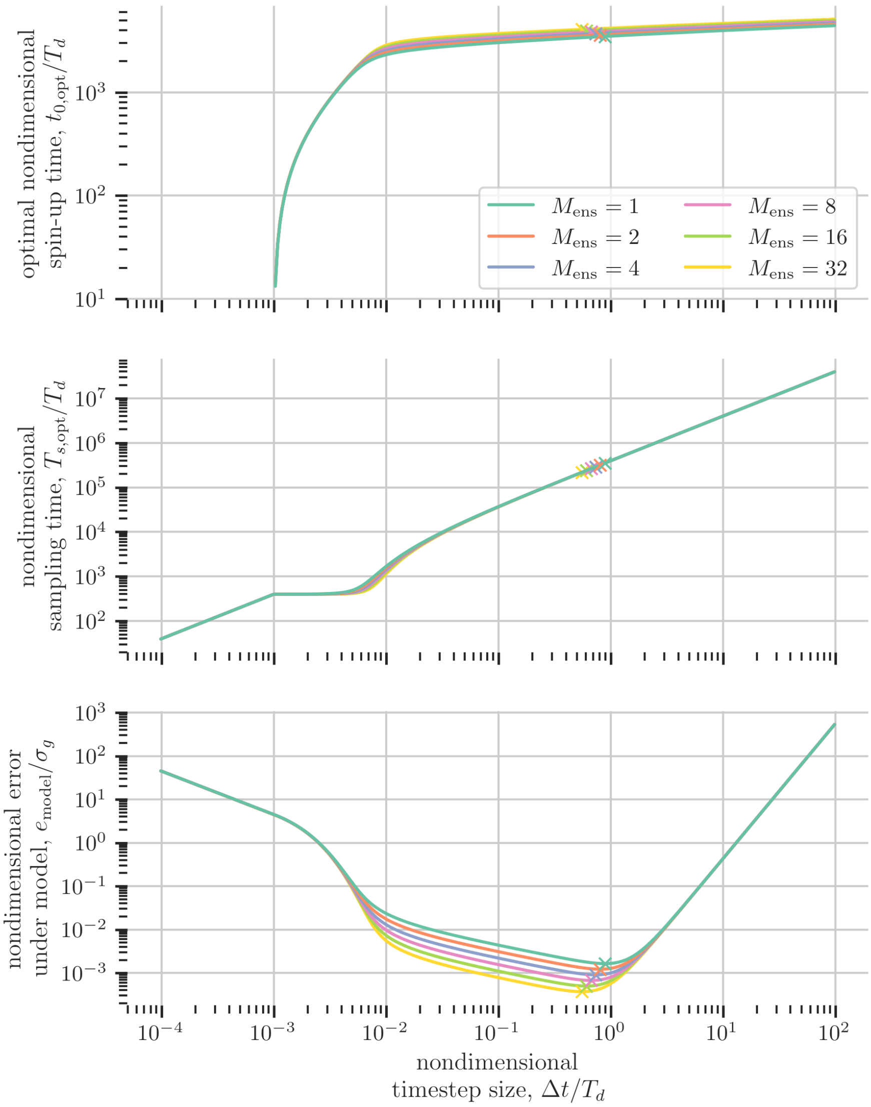


Figure 1.19: Dependence of normalized spin-up time t_0/T_d, sampling time T_s/T_d, and model error $e_{\text{model}}/\sigma_g$ on normalized timestep $\Delta t/T_d$ with total cost set at $U = 1.2 \times 10^6$ for 3rd-order Runge Kutta.

The bottom plot of Figure 1.19 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 (1.53) is controlled shifts from the $\exp(-t_0)$

term to the T_s^{-1} term as $\Delta t / T_d \to 0$, 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 $e_{\mathrm{model,MC}}$ with U under the optimal choices and evaluate how well $e_{\mathrm{model,MC}}$ approximates experimental data for $\mathbb{E}[|J_{\mathrm{MC}} - J_\infty|]$. In Figure 1.20, the variation of $e_{\mathrm{model,MC}}$ computed via (1.64) 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 (1.44), 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 (1.19). 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.

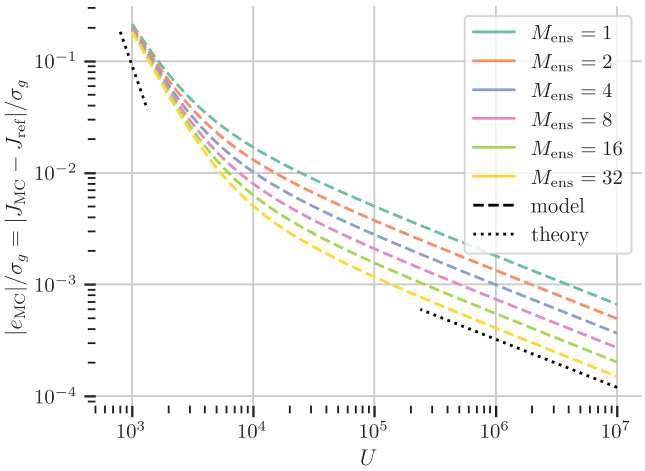


Figure 1.20: Optimal non-dimensional error under model as a function of total cost U for RK3. Theory totem on left-hand side: discrete convergence rate, $1/q$; on right-hand side: $\frac{2(q+r)}{q}$ rate from (1.44).

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 $(\Delta t)_{\mathrm{MC,opt}}$ and $N_{\mathrm{MC,opt}}$ and using the model fit given in Table 1.2. In Figures 1.21, 1.22, and 1.23, we show the predictions and the results of Monte Carlo estimates of $\mathbb{E}[|J_{\mathrm{MC}} - J_\infty|]$ 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.

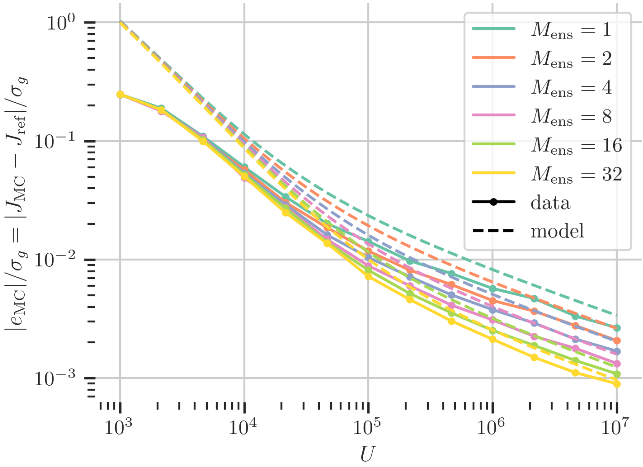


Figure 1.21: Total cost model and Monte Carlo validation as a function of total cost U for FE.

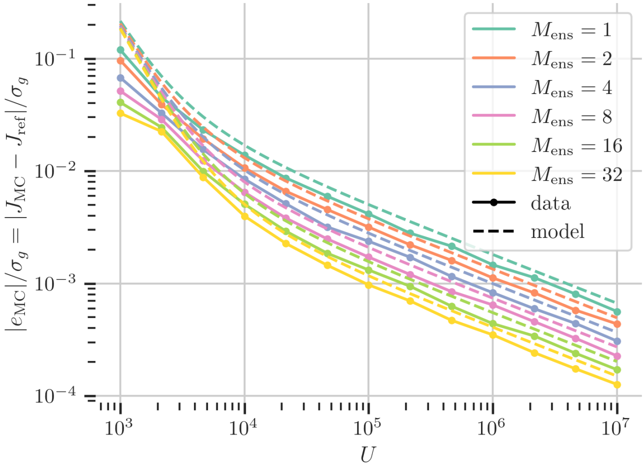


Figure 1.22: Total cost model and Monte Carlo validation as a function of total cost U for RK3.

1.7 *Conclusion*

In this chapter, 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 outstanding problem is the expense of identifying the parameters of the error model. In order to overcome this, we will show in the next chapter that leveraging a Bayesian approach, as in Oliver

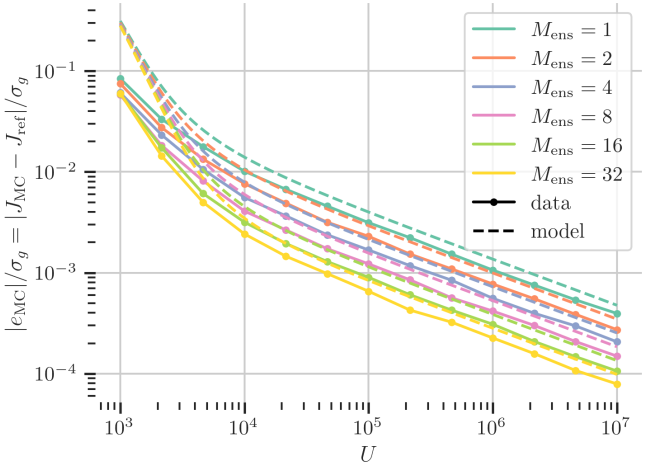


Figure 1.23: Total cost model and Monte Carlo validation as a function of total cost U for RK4.

et al. [2014], can allow us to approximate the model in (1.20) at relatively small cost. Then in the proceeding chapter, we will need to extend the framework to 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 *and* spatial discretization and act optimally with respect to both. Last but not least, we can exploit the results to conduct a high-fidelity simulation at an (approximately) optimal discretization.

2

Bayesian small sample identification of error models for discretizations of ordinary differential equations

You can observe a lot just by watching.

—Yogi Berra

To effectively use error models for chaotic systems, we must be able to know them before setting up and starting a simulation. In Chapter 1, we showed that the error in approximating the true mean value J_∞ of a ergodic chaotic ODE system could be well approximated by an error model of the form

$$e_{\text{model}} = e_{\text{model}}(\Delta t, t_0, T_s;\ M_{\text{ens}}, \theta, \psi)$$

given in the most general form in (1.58). Here, $\psi = (|A_\lambda|, T_\lambda)$ and $\theta = (C_q, q, A_0, r)$ give parametrizations of the transient error and attractor error models, respectively. In Section 1.5, a method for approximating ψ using the discrete output g_{hp} of a chaotic system.

In this chapter, we concentrate on identifying the parameters θ, operating with $t_0 \gg T_\lambda$ such that e_λ is negligible and the error model reduces to:

$$\begin{aligned} e_{\text{model}} &= e_{\text{model}}(\Delta t, T_s; \theta) \\ &= C_q \Delta t^q + \frac{A_0}{\sqrt{M_{\text{ens}}}} T_s^{-r}, \end{aligned} \tag{2.1}$$

This model implies an optimal choice of Δt_{opt} and $T_{s,\text{opt}} = N_s \Delta t_{\text{opt}}$ at a given number of sampling timesteps N_s:

$$\Delta t_{\text{opt}}(N_s;\,\theta) = M_{\text{ens}}^{-\frac{1}{2(q+r)}} \left(\frac{rA_0}{qC_q}\right)^{\frac{1}{q+r}} N_s^{-\frac{r}{q+r}} \tag{2.2}$$

$$T_{s,\text{opt}}(N_s;\,\theta) = M_{\text{ens}}^{-\frac{1}{2(q+r)}} \left(\frac{rA_0}{qC_q}\right)^{\frac{1}{q+r}} N_s^{\frac{q}{q+r}}, \tag{2.3}$$

at which the optimal (i.e. minimal) absolute error in expectation would be achieved:

$$e_{\text{opt}}(N_s;\theta) = M_{\text{ens}}^{\frac{q}{2(q+r)}} A_0^{\frac{q}{q+r}} C_q^{\frac{r}{q+r}} \left(\left(\frac{r}{q}\right)^{\frac{q}{q+r}} + \left(\frac{r}{q}\right)^{-\frac{r}{q+r}}\right) N_s^{-\frac{qr}{q+r}}. \tag{2.4}$$

We note that (2.2), (2.3), and (2.4) are given here in the equivalent dimensional form to those in Chapter 1, in which $r \to 1/2$ is assumed a priori. Thus, given θ, optimal choices for Δt and T_s can be made to achieve the minimum expected error in $J_{T,hp}$. In our previous work, we showed we could fit (2.1) to identify θ using expensive Monte Carlo estimates of $\mathbb{E}[|e_{T,hp}|]$ that we generated after computing an additional very high cost approximation J_{ref} of J_∞. However, in practice, expensive Monte Carlo studies and reference values will not be available, so we set out in this work to estimate the model in (2.1) with a small number of simulations of $J_{T,hp}$.

In Oliver et al. [2014], the authors develop an auto-regressive (AR) modeling approach to approximate the statistical error behavior that underlies the coefficient A_0 in (2.1) and then apply this approximation to create a Bayesian method for Richardson extrapolation, which effectively then allows the estimation of C_q and J_∞. Unfortunately, AR modeling techniques tend to suffer from noisy and irregular approximation when applied to signals from real systems, because the correlations in physical processes can be more complicated than the assumed auto-regressive model; this is reflected in Oliver et al. [2014], as the authors must develop a complicated model selection procedure to arrive at their estimates, and they report frequent user intervention in the presence of AR model instabilities and errors.

An alternative strategy is using statistical methods for outright identification of (2.1). Success has been demonstrated in balancing discretization and stochastic errors in steady simulations of heterogeneous media using frequentist statistical approaches incorporated into multi-level Monte Carlo schemes[1]. A framework like this one is promising for chaotic systems, where unpredictability replaces stochasticity. However, it requires significant sample sizes in order to make approximations of the error contributions at a given budget

[1] Florian Müller, Patrick Jenny, and Daniel W. Meyer. Parallel multilevel Monte Carlo for two-phase flow and transport in random heterogeneous porous media with sampling-error and discretization-error balancing. *SPE Journal*, 21(06):2027–2037, 09 2016

level and, therefore, to balance the error contributions at the next level of a simulation. On the other hand, the challenge posed by the results in Chapter 1 is achieving identification or approximation of the error models without massive numbers of high-cost simulations.

In this chapter we show that a small sample Bayesian approach can effectively approximate the terms in (2.1) without expensive reference computations. We demonstrate that the asymptotic discretization and sampling error models can be modified to generate a likelihood function that describe the outputs $J_{T,hp}$ of any given simulation given the true solution and parameters related to θ. Then, we employ this likelihood in a Bayesian method that allows reliable approximation of the model using a small sample of low-cost $J_{T,hp}$ simulations and enables the selection of Δt and T_s that are optimal under the model.

2.1 *Bayesian error modeling*

Bayesian likelihood formulation

Following the previous work, we break down the error incurred by the dual approximations in $J_\infty \approx J_{T,hp}$. We follow Section 1.1 up to (1.7), which is:

$$e_{T,hp} = \underbrace{(J_{T,hp} - J_T)}_{e_{hp}} + \underbrace{(J_T - J_\infty)}_{e_T} \tag{1.7}$$

Now, we can insert (1.5) into (1.7) and rearrange to find

$$J_{T,hp} = J_\infty + e_{hp} + e_T. \tag{2.5}$$

We assume that the discretization error has the form:

$$e_{hp} \approx C_q^*(\Delta t)^q, \tag{2.6}$$

with $q \in \mathbb{R}^+$ and $C_q^* \in \mathbb{R}$.

For the statistical error term, we now revisit the central limit theorem (CLT) for dynamical systems[2]. Under the asymptotic behavior of the CLT we can expect the quasi-random effect of sampling to take the form of a normal random variable:

$$e_T \sim \mathcal{N}\left(0, \left(\frac{A_0^*}{\sqrt{M_{\text{ens}}}} T_s^{-r}\right)^2\right), \tag{2.7}$$

with $A_0^*, r \in \mathbb{R}^+$, $r \to 1/2$ as $T_s \to \infty$. If we take (2.6) and (2.7) and insert them into (2.5), we arrive at:

$$J_{T,hp} \sim \mathcal{N}\left(J_\infty + C_q^*(\Delta t)^q, \left(\frac{A_0^*}{\sqrt{M_{\text{ens}}}} T_s^{-r}\right)^2\right), \tag{2.8}$$

[2] Denker, 1989

which allows us to describe the output quantity of the discrete system as a random variable, given knowledge of the augmented parameter set

$$\theta^* = \left(C_q^*, q, A_0^*, r, J_\infty\right). \tag{2.9}$$

Manipulating (2.8), we can find:

$$\begin{aligned}
\mathbb{E}[|J_{T,hp} - J_\infty|] &= \mathbb{E}\left[\left|\mathcal{N}\left(C_q^*(\Delta t)^q, \left(\frac{A_0^*}{\sqrt{M_{\text{ens}}}} T_s^{-r}\right)^2\right)\right|\right] \\
&= \mathbb{E}\left[\left|C_q^*(\Delta t)^q + \mathcal{N}\left(0, \left(\frac{A_0^*}{\sqrt{M_{\text{ens}}}} T_s^{-r}\right)^2\right)\right|\right] \\
&\leq \mathbb{E}\left[\left|C_q^*(\Delta t)^q\right|\right] + \mathbb{E}\left[\left|\mathcal{N}\left(0, \left(\frac{A_0^*}{\sqrt{M_{\text{ens}}}} T_s^{-r}\right)^2\right)\right|\right] \\
&\leq |C_q^*|\ (\Delta t)^q + \sqrt{\frac{2}{\pi}} \frac{A_0^*}{\sqrt{M_{\text{ens}}}} T_s^{-r}.
\end{aligned} \tag{2.10}$$

Thus, the likelihood formulation and its parameters map to a model of the same form as $e_{\text{model}} = e_{\text{model}}(\,\cdot\,;\theta)$ in (2.1) under the transformation:

$$C_q = |C_q^*| \qquad A_0 = \sqrt{\frac{2}{\pi}} A_0^*. \tag{2.11}$$

Alternately, we can write:

$$\begin{aligned}
\mathbb{E}[|J_{T,hp} - J_\infty|] &= \mathbb{E}\left[\left|\mathcal{N}\left(C_q^*(\Delta t)^q, \left(\frac{A_0^*}{\sqrt{M_{\text{ens}}}} T_s^{-r}\right)^2\right)\right|\right] \\
&= \mathbb{E}\left[\mathcal{F}\left(C_q^*(\Delta t)^q, \left(\frac{A_0^*}{\sqrt{M_{\text{ens}}}} T_s^{-r}\right)^2\right)\right].
\end{aligned} \tag{2.12}$$

where $\mathcal{F}(\mu, \sigma^2)$ gives a folded normal distribution with a mean μ and standard deviation σ, such that $|X| \sim \mathcal{F}(\mu, \sigma^2)$ when $X \sim \mathcal{N}(\mu, \sigma^2)$. We can write the expectation of a folded normal distribution with an underlying mean μ and standard deviation σ as

$$\mathbb{E}[\mathcal{F}(\mu, \sigma)] = \sigma\sqrt{\frac{2}{\pi}} \exp\left(-\frac{\mu^2}{2\sigma^2}\right) + |\mu| \operatorname{erf}\left(\frac{|\mu|}{\sqrt{2\sigma^2}}\right), \tag{2.13}$$

(noting that $\operatorname{sign}(|x| \operatorname{erf}(|x|/C)) = \operatorname{sign}(x \operatorname{erf}(x/C))$ will hold $\forall C > 0$). Thus, substituting $\mu = C_q^* \Delta t^q$ and $\sigma = A_0^* M_{\text{ens}}^{-1/2} T_s^{-r} = A_0^* M_{\text{ens}}^{-1/2} N_s^{-r} \Delta t^{-r}$ gives an alternative "folded" error model. In the forthcoming results, we will show that the error model based on the folded normal output model will slightly more accurately describe the output error behavior– in the sense that 50% of results should be above and below the expected error model– but at the cost of analytical complexity. On the other hand, the original error model will give an upper bound on the folded error model and offer more analytical insight. For this reason, "optimal" results in this work will refer to optimality with respect to (2.1) under the transformation in (2.11).

Finally, using (2.8) the likelihood of $J_{T,hp}$ conditioned on θ^* can be explicitly written as

$$p\left(J_{T,hp} \;\middle|\; \theta^*\right) = \mathcal{N}\left(J_{T,hp};\, J_\infty + C_q^*(\Delta t)^q, \left(\frac{A_0^*}{\sqrt{M_\text{ens}}} T_s^{-r}\right)^2\right), \quad (2.14)$$

which can be exploited by Bayesian methods. Here, the notation $\mathcal{N}(x;\mu,\sigma^2)$ refers to the probability of an event $X = x$ where the random variable $X \sim \mathcal{N}(\mu,\sigma^2)$. If we use independent initial conditions for each simulation and spin up each until the solution is on the attractor, then each of the discrete results of the chaotic system will be independent random samples from a stationary distribution. We can thus write the conditional likelihood of a set of M simulations $\{J_{T,hp}\}$:

$$p\left(\{J_{T,hp}\} \;\middle|\; \theta^*\right) = \prod_{m=1}^{M} p\left(J_{T,hp}^{(m)} \;\middle|\; \theta^*\right). \quad (2.15)$$

Then, using Bayes theorem, the likelihood of the set of parameters θ^* conditioned on the results $\left\{J_{T,hp}\right\}$ is:

$$p\left(\theta^* \;\middle|\; \{J_{T,hp}\}\right) = \frac{p\left(\{J_{T,hp}\} \;\middle|\; \theta^*\right) p\left(\theta^*\right)}{p\left(\{J_{T,hp}\}\right)}. \quad (2.16)$$

Since the marginal likelihood of the set of output data $p\left(\{J_{T,hp}\}\right)$ is a constant for any given set of output data, we can therefore design estimation methods to estimate $p\left(\theta^* \;\middle|\; \{J_{T,hp}\}\right)$ up to a constant factor without the need to account for $p\left(\{J_{T,hp}\}\right)$ directly. Finally, we will need to treat the prior likelihood of the data $p\left(\theta^*\right)$.

Prior model formulation

In Bayesian inference, a prior model for $p(\theta^*)$ is assumed; by the Bernstein-von Mises theorem, it can be shown that as the number of datapoints goes to infinity, the posterior estimate from a well posed likelihood and prior model $p\left(\theta^* \;\middle|\; \{J_{T,hp}\}\right)$ will converge towards the same result as the maximum likelihood estimator independently of the choice of prior[3].

[3] David A. Freedman. On the asymptotic behavior of Bayes' estimates in the discrete case. *The Annals of Mathematical Statistics*, 34(4):1386–1403, 12 1963

For the problem at hand, we now discuss building a prior for θ^*. We begin with q and r; for these, by the appropriate truncation error analysis and the central limit theorem, we can determine theoretical asymptotic values of q and r as $\Delta t \rightarrow 0$ and $T_s \rightarrow \infty$, respectively. Of course, at finite Δt and T_s, the rates are often not exactly the theoretical asymptotic values. From experience running discrete simulations near asymptotic convergence, we guess q to be approximately within bounds

$$q_{\text{theory}} - \frac{1}{2} \lesssim q_{\text{guess}} \lesssim q_{\text{theory}} + \frac{1}{2}$$

and similarly for r

$$\frac{1}{2} - \frac{1}{16} \lesssim r_{\text{guess}} \lesssim \frac{1}{2} + \frac{1}{16}.$$

In the forthcoming work of this chapter, we use the RK3 methods as described in Section 1.2; thus, $q_{\text{theory}} = 3$ is the derived result for RK3 from numerical analysis, assuming that $\mathbf{u}$ is sufficiently smooth. Meanwhile, $r_{\text{theory}} = 1/2$ using the asymptotic rate of the CLT. In practice, the convergence rate of the output may vary depending on the system and the choice of g, and r might deviate from $1/2$ if the CLT does not apply, due to uncontrolled transient effects or insufficient T_s. We can take this qualitative description to map to a prior distribution on q and r. We want a prior distribution that has support on the positive real numbers with a well-defined mean and standard deviation; to achieve this, we choose a gamma distribution and interpret the bounds on q_{guess} and r_{guess} as 95% bounds. Figure 2.1 shows the prior distributions for the present case, a Runge-Kutta 3$^{\text{rd}}$-order discretization of the Lorenz system.

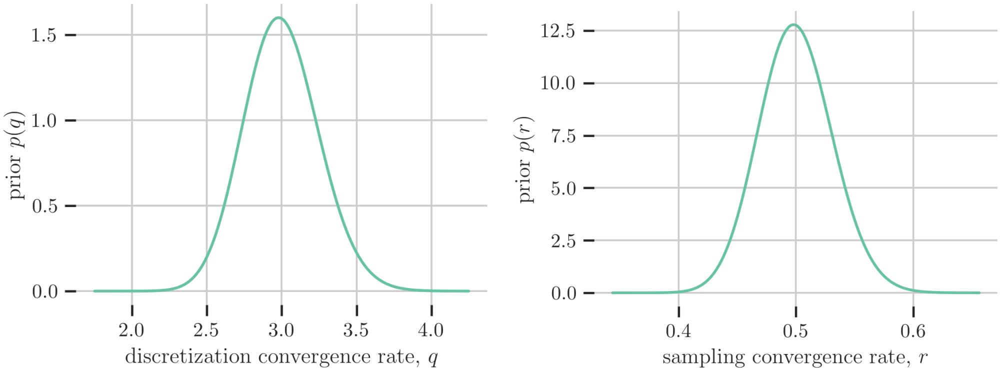


Figure 2.1: Prior distributions for convergence rate variables in the model.

The next prior to define is the true value of the output, which is unknown in general. We often have some experience to estimate the value of J_∞ which is less theoretically-based than for the rates q and r

but nonetheless gives a range of plausible values of J_∞. As concerns the present case, discrete traces of the Lorenz system, such as the output trace in Figure 2.2, are prevalent in the literature. Using these types of results, we can make an estimate for the output as

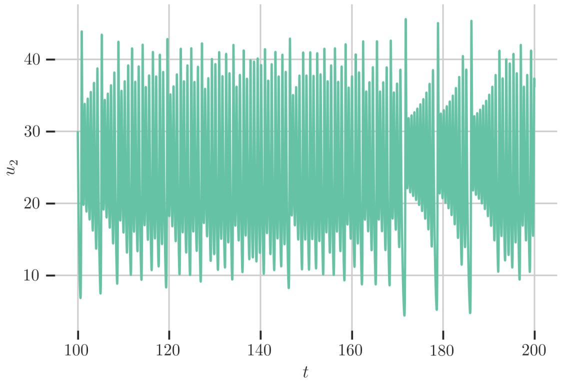


Figure 2.2: Instantaneous output trace of $g = u_2$ of discrete Lorenz system with $\Delta t = 10^{-3}$ RK4 discretization.

These types of prior distributions represent the injection of the simulation designer's intuition. In practice, this would look like making ballpark estimates, like an aerodynamicist estimating 0.05 ± 0.05 for a drag coefficient, or using a statistical analysis on historical data to project estimates of a wind farm's output.

$$J_{\text{guess}} = 23 \pm 1. \tag{2.17}$$

In general, outputs are real numbers, can be positive or negative, and will have bounded statistics. Thus, we reinterpret this quantitatively as a 95% confidence interval on a normal random variable, and in Figure 2.3, we can see the prior distribution that it implies.

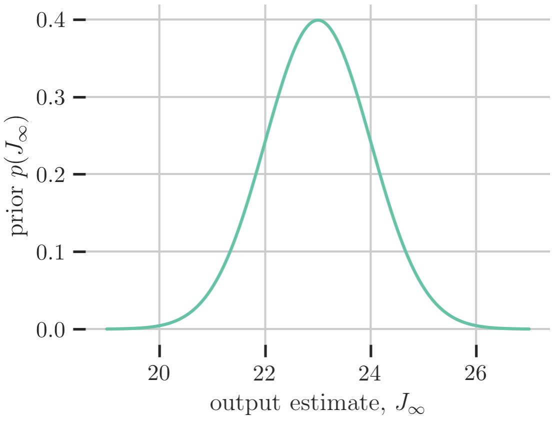


Figure 2.3: Prior distribution for J_∞ estimate.

Finally, we must specify values for C_q^* and A_0^*. Often, the goal of a computational study of a chaotic system is not solely to consider one particular set of parameters α, but rather over a range of possible values. In this case, prior models for C_q^* and A_0^* can be developed using this previous knowledge; however, even in this case making estimates beyond order-of-magnitude estimates relies strongly on the assumption that J_∞ will vary smoothly with α, which is not assured with nonlinear dynamical systems. Conservatively, we will want to make guesses about the maximum order of magnitude of these quantities. In the specific case of the Lorenz system study here, we can refer to the values of C_q and A_0 found in Chapter 1 and given in Table 1.2. In order to emulate the increased uncertainty about these values that we might expect in practice, we choose C_{guess} and A_{guess} to have three orders of magnitude more uncertainty in the values than we observe in the reference values, resulting in:

$$-1 \times 10^6 \lesssim C_{\text{guess}} \lesssim 1 \times 10^6$$

and

$$0 \lesssim A_{\text{guess}} \lesssim 2 \times 10^3.$$

In general, we now want to choose a prior distribution that has values of about the same order of magnitude for all reasonable C_q^* and A_0^*. At the same time, the prior distributions should have support on all possible values, in order to satisfy the Bernstein-von Mises theorem. We can achieve this with a normal distribution for C_q^* and a half-normal distribution for A_0^*. As previously, we translate the C_{guess} and A_{guess} ranges into quantitative ones by taking them as 95% confidence intervals. In Figure 2.4, we can show these prior distributions for the example of a Runge-Kutta 3rd-order discretization of the Lorenz system. We can see that one deficiency of these choices is that $C_q^* = A_0^* = 0$ maximize the prior model, though the expectation of $|C_q^*|$ and A_0^* are non-zero.

A common alternative non-negative distribution would be the lognormal distribution, for example, but it is not preferable because it implies a preference for a particular range of magnitudes, whereas all values below the σ parameter for a halfnormal variable have a similar likelihood.

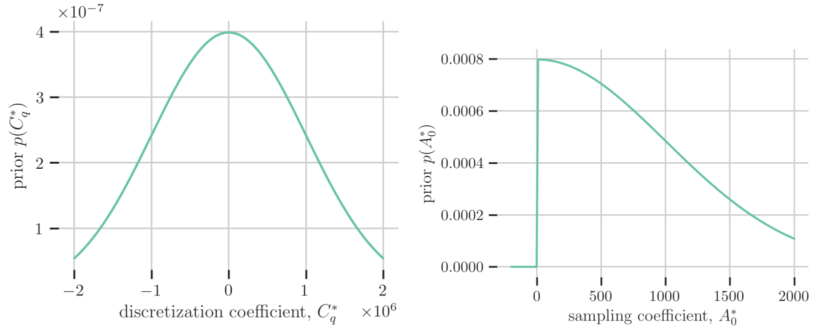


Figure 2.4: Prior distributions for constant factor variables in the model.

The prior distributions that we arrive at, then, are given by:

$$\begin{aligned} C_q^* &\sim \mathcal{N}(\mu_C, \sigma_C^2), \\ q &\sim \Gamma(\alpha_q, \beta_q), \\ A_0^* &\sim \mathcal{H}(\sigma_A^2), \\ r &\sim \Gamma(\alpha_r, \beta_r), \text{ and} \\ J_\infty &\sim \mathcal{N}(\mu_J, \sigma_J^2), \end{aligned} \tag{2.18}$$

The resulting prior model parameters are given in Table 2.1.

With these five distributions, $p(q)$, $p(r)$, $p(C_q^*)$, $p(A_0^*)$, and $p(J_\infty)$, we can now combine them, taking each to be independent, to arrive at a fully specified prior:

$$p\left(\theta^*\right) = p(C_q^*)\, p(q)\, p(A_0^*)\, p(r)\, p(J_\infty). \tag{2.19}$$

Taking (2.19) and substituting into (2.16), we can now write the posterior distribution up to a constant factor. This allows us to evaluate the maximum a posteriori (MAP) estimator of θ^* using an appropriate optimization method, and enables the use of Markov Chain Monte Carlo (MCMC) methods to draw samples from the posterior distribution.

	q	r
μ	3.0	0.5
σ	0.25	0.031,25
α	144.0	256.0
β	48.0	512.0

	C_q^*	J_∞
μ	0.0	23.0
σ	1,000,000	0.5

	A_0^*
σ	10^3

Table 2.1: Parameters for prior specification for RK3 error model. α and β variables are the parameters for Gamma distributions, μ and σ represent mean and standard deviations observed (q and r) or specified (C_q^*, A_0^*, J_∞).

2.2 *Numerical results*

Bayesian fit results

Using the numerical methods and problem setup outlined in Section 1.2, we have collected a library of 1×10^5 samples of $J_{T,hp}$, at $N_s = 1 \times 10^5$. For simplicity, we validate using $M_{\text{ens}} = 1$ only. For any given sample $J_{T,hp}^{(i)}$, we use a choice of $\Delta t^{(i)}$ that is sampled randomly, such that:

$$\log_{10}(\Delta t) \sim \mathcal{U}(\log_{10}(\Delta t_{\text{min}}), \log_{10}(\Delta t_{\text{max}})), \tag{2.20}$$

where $\mathcal{U}$ the uniform distribution, and where we use $T_{s,\text{min}} = 10$ and $\Delta t_{\text{max}} = 3 \times 10^{-2}$ in order to attempt to eliminate the non-asymptotic behavior for small T_s and large Δt, respectively, setting $\Delta t_{\text{min}} = T_{s,\text{min}}/N_s$. Now, we will take samples out of this library length-M sets $\{J_{T,hp}\}$, against which we calculate estimates of θ^*.

Now, we will evaluate the posterior estimates of θ^* and the error models that they imply. The result of the Bayesian inference problem is not a unique single solution for the error model, but a random variable that describes the posterior likelihood of any given parameter set θ^*. To begin, we will assess the resulting posterior distributions by sampling to evaluate their spread. In order to sample out of the posterior distributions, we have implemented (2.16), (2.18), and (2.19) in the `Stan` language[4]. `Stan` implements Hamiltonian Monte Carlo (HMC) using a state-of-the-art No U-Turn Sampler, which provides samples out of the posterior distribution in computation time negligible compared to the cost generating the posterior samples. In our work, all HMC results use $H = 4$ chains, each with $2S = 10 \times 10^3$ total samples, where the first S samples from each

[4] Stan Developer Team. Stan modeling language users guide and reference manual. Technical report, 2021

chain are discarded "burn-in", which is used in order to arrive at the stationary distribution before sampling. This leaves $H \times S$ total samples from the posterior.

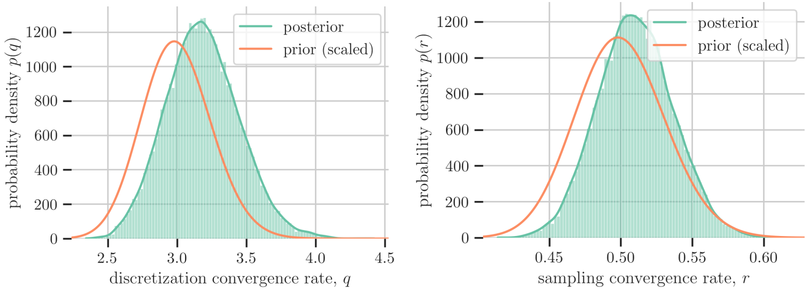


Figure 2.5: Comparison between prior and posterior distributions for rate parameters q and r for $M = 100$ 3rd-order Runge-Kutta discretizations of the Lorenz system using $N_s = 1 \times 10^5$ timesteps each.

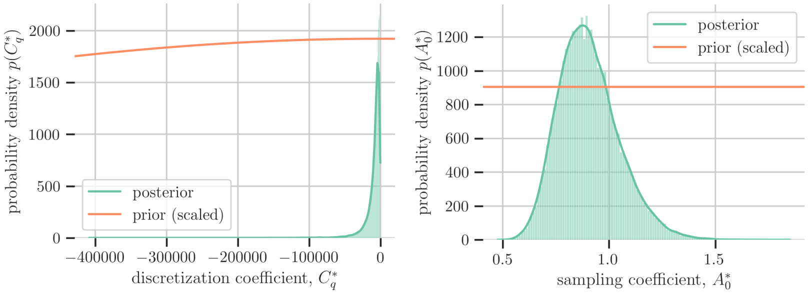


Figure 2.6: Comparison between prior and posterior distributions for leading coefficients C_q^* and A_0^* for $M = 100$ 3rd-order Runge-Kutta discretizations of the Lorenz system using $N_s = 1 \times 10^5$ timesteps each.

In Figures 2.5 and 2.6, we examine histograms of each posterior parameter using the HMC samples for one dataset of random $M = 100$ simulations, overlaid with the prior models for each. From Figure 2.5, we can see that the posterior estimate of the rates vary away from the prior model rates, while the order of magnitude of uncertainties is similar. Figure 2.6, on the other hand, shows significant decreases in uncertainty on the wide priors for C_q^* and A_0^*. Last, the posterior estimate and prior assumption for J_∞ are shown in Figure 2.7. Here, again, significant improvement in the uncertainty around the output estimate is demonstrated.

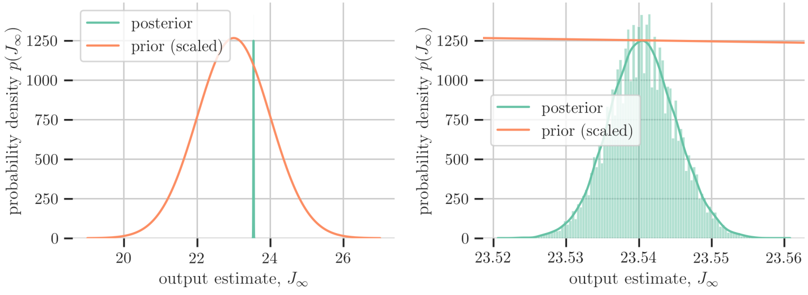


Figure 2.7: Comparison between prior and posterior distributions for identification of J_∞ using $M = 100$ 3rd-order Runge-Kutta discretizations of the Lorenz system using $N_s = 1 \times 10^5$ timesteps each.

The above posterior estimates represent marginal distributions of a five-dimensional posterior distribution on θ^*. In general, the components of θ^* on the posterior are not independent, but will have some correlations between them. In Figure 2.8, the full five-dimensional distribution of θ^* is visualized using cross-correlations for one selection of $M = 100$ discrete results, showing the (sometimes) complex interactions between the components of θ^*.

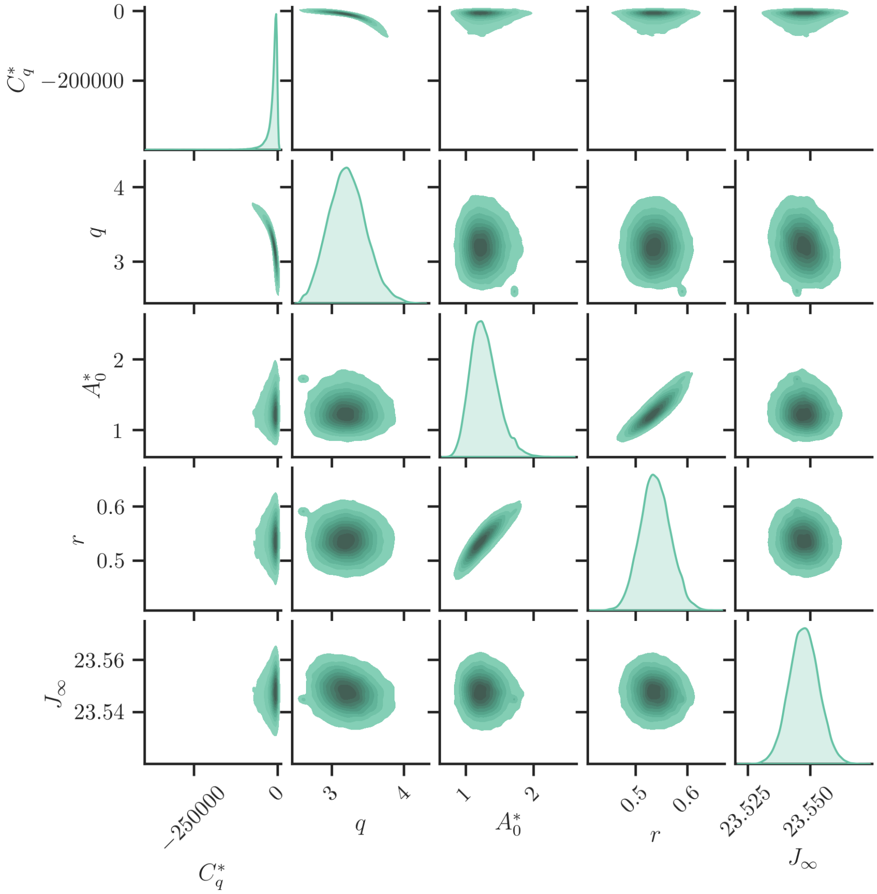


Figure 2.8: Cross correlation of Hamiltonian Monte Carlo samples from posterior distribution for $M = 100$

In addition to sampling with the HMC, we also compute maximum a posteriori (MAP) estimates, which allow us to find the most probable θ^* on the full five-dimensional posterior distribution, accounting for their inter-dependencies:

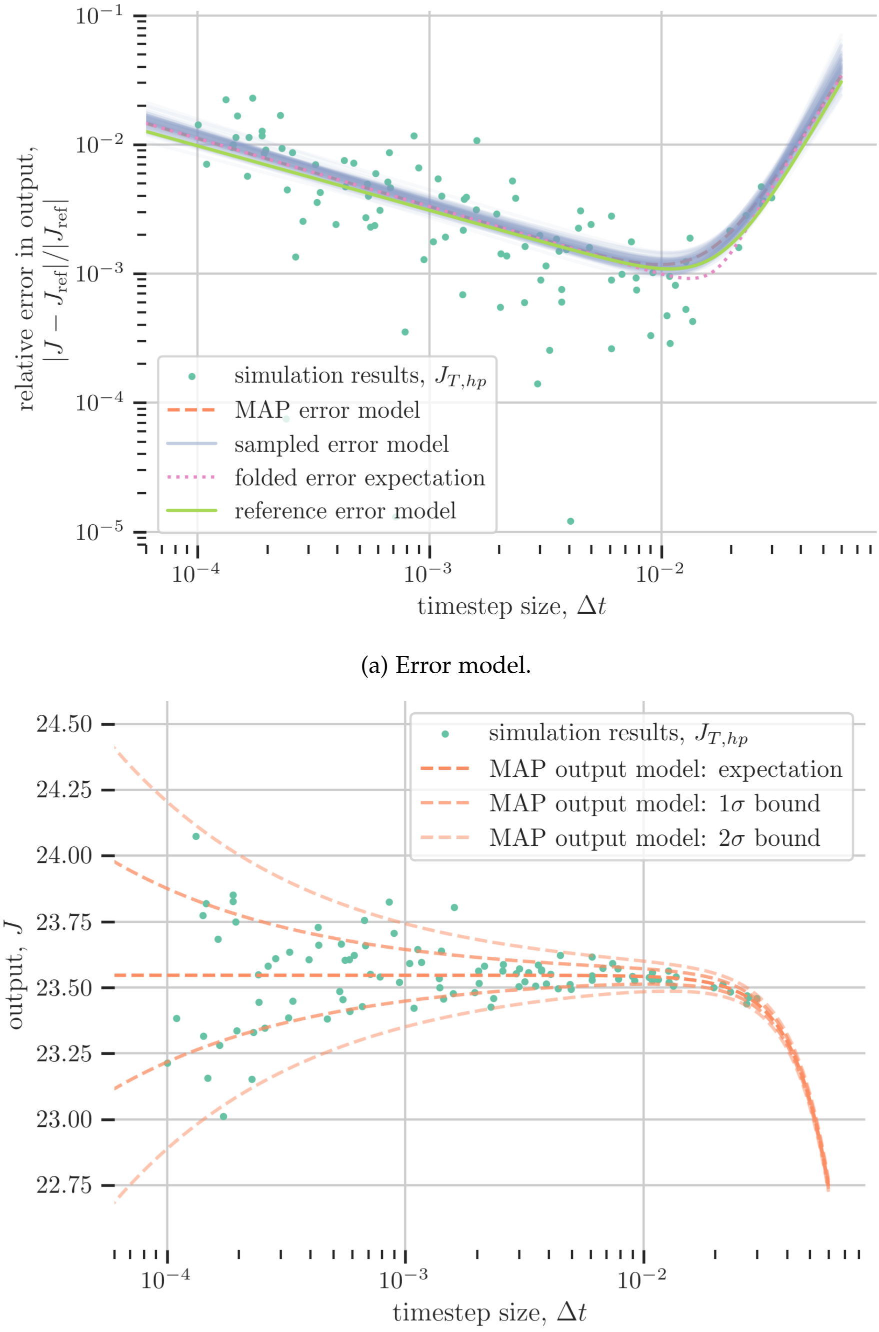
$$\theta^*_{\text{MAP}} = \max_{\theta^*} \log p\left(\theta^* \,\middle|\, \{J_{T,hp}\}\right), \tag{2.21}$$

where $p\left(\theta^* \,\middle|\, \{J_{T,hp}\}\right)$ is the posterior likelihood given in (2.16). Like in the HMC samples, θ^*_{MAP} is computed using the optimization schemes in `Stan`.

Now, we want to evaluate the error models generated with the

Bayesian inference method. We use the reference fit in (1.23) and the reference model in Table 1.2 for comparison. In Figure 2.9a, we show the error model estimates developed by the small-sample method. The plot shows the MAP error model and models associated with each of 100 HMC samples from the posterior as well as the folded error expectation given by supplying (2.13) with θ^*_{MAP}. These are plotted with the high-cost nonlinear least squares (NLS) reference model, $e_{\mathrm{model,ref}}$, from Table 1.2, with which all of the error models demonstrate strong agreement. In this plot, we see that the Bayesian process broadly captures the behavior of the error model using this particular $M = 100$ set, $\{J_{T,hp}\}$. It accurately predicts where Δt_{opt} will occur, and its guess of e_{opt} is off by a small factor, less than 2. To further understand the behavior of the fit, we consider the likelihood model for the output, (2.14), from which the error model estimates are derived, shown in Figure 2.9b. Here, the expected output is plotted with a dark orange dashed line, and the one and two standard deviation bounds from the likelihood model are plotted with subsequently lighter dashed lines. We can see here that the behavior of the output $J_{T,hp}$ matches qualitatively with expectations. When Δt and $T_s = N_s \Delta t$ are small, simulations vary significantly about a nearly constant mean value, which approaches J_∞. For large Δt and T_s, random variation grows small about a growing mean offset.

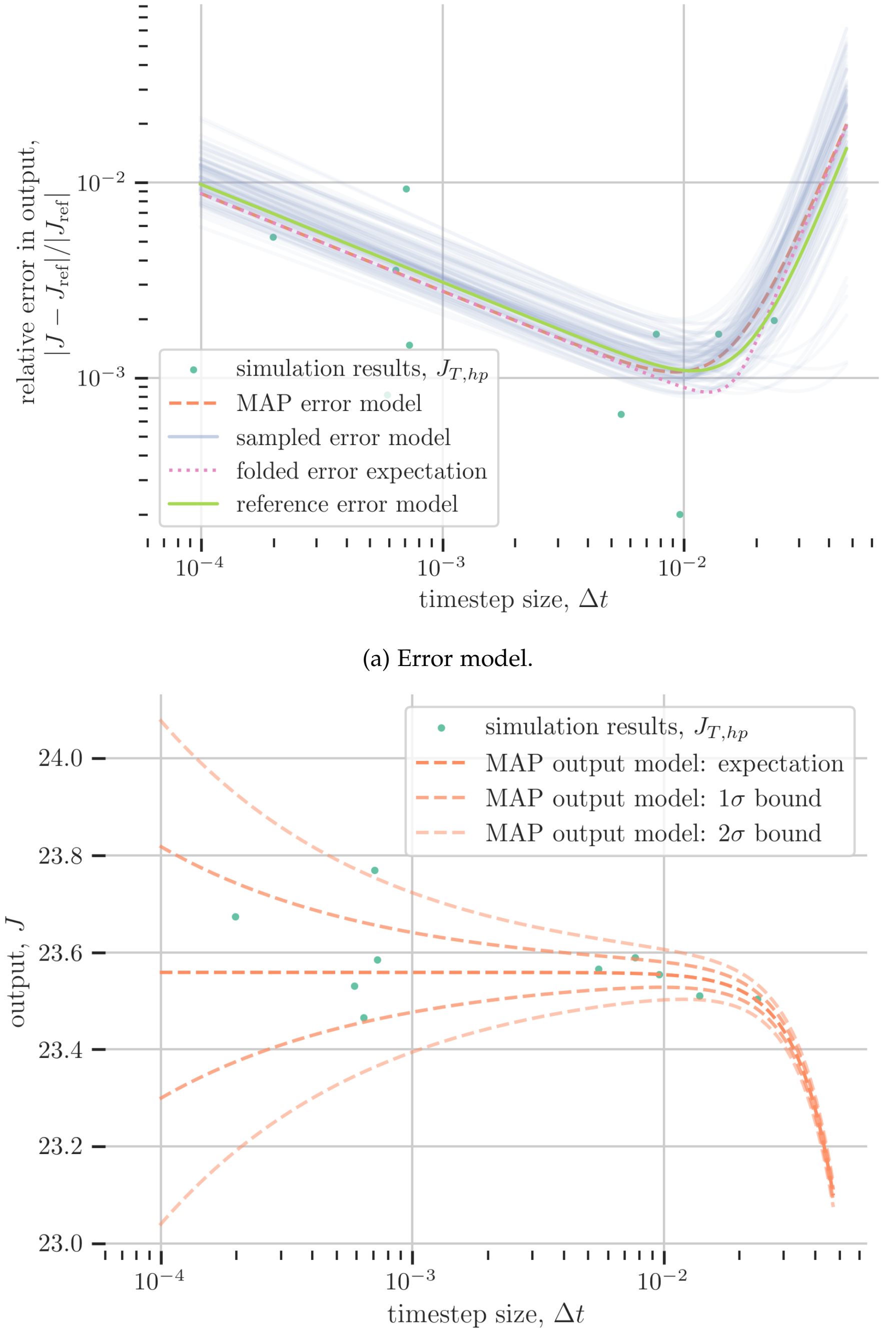
We can repeat this process identically for only ten randomly drawn simulations, with no other changes. We omit the posterior studies this time, and the resulting error model and likelihood model are shown in Figure 2.10. Immediately, we can see that the spread in the sampled error models is significantly larger than in Figure 2.9. While there is more uncertainty, we can also see that the Bayesian informative prior approach that we've outlined in this work is able to regularize the small sample inference problem such that we can identify the error model. Moreover, we do so in a way that matches closely with the result of the high-cost structured NLS fit from the previous work, for at least this particular random selection of $M = 10$ simulations. In a Section 2.2 we will consider the behavior of this method in expectation across many possible length-M sets $\{J_{T,hp}\}$. Before doing so, we repeat the fit process using a larger set of $M = 1000$ simulations. Figure 2.11 gives the results of this simulation, which demonstrates very close agreement with the reference error model and an even tighter spread of sampled error models.



(a) Error model.

(b) Output model.

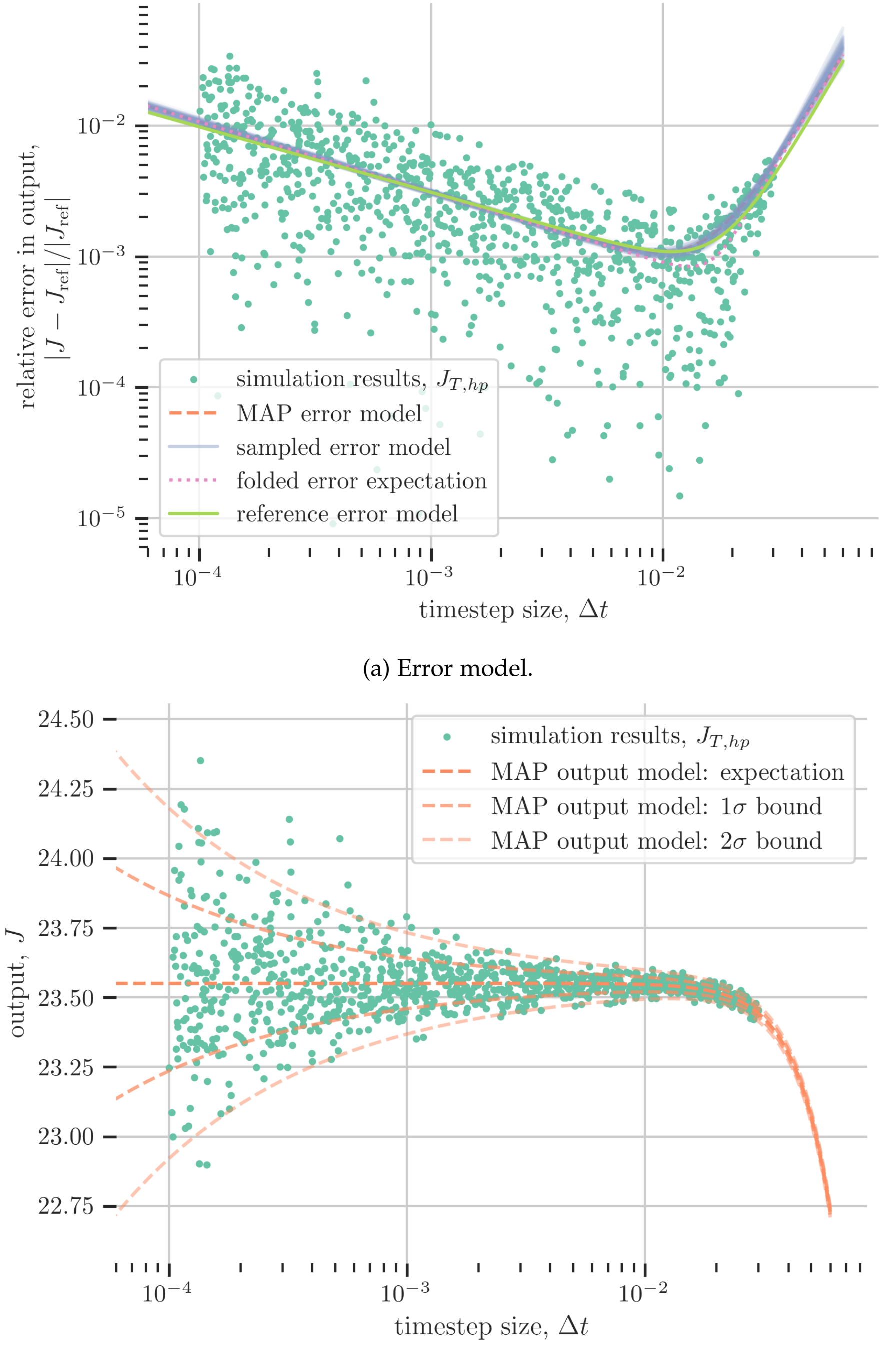
Figure 2.9: Bayesian approximation of models for $M = 100$ 3rd-order Runge-Kutta discretizations of the Lorenz system using $N_s = 1 \times 10^5$ timesteps each.



(a) Error model.

(b) Output model.

Figure 2.10: Bayesian approximation of models for $M = 10$ 3rd-order Runge-Kutta discretizations of the Lorenz system using $N_s = 1 \times 10^5$ timesteps each.



(a) Error model.

(b) Output model.

Figure 2.11: Bayesian approximation of models for $M = 1000$ 3rd-order Runge-Kutta discretizations of the Lorenz system using $N_s = 1 \times 10^5$ timesteps each.

Performance in expectation as a function of M

In order to understand the variability of the Bayesian procedure outlined in this section, we now perform Monte Carlo simulations for a range of M. We will bootstrap length M sets out of the library of simulations detailed in Section 2.2 then find the MAP estimate θ^*_MAP from each length-M set, $\{J_{T,hp}\}$. This is repeated over an ensemble of size $\mathcal{M} = 100$ at each M. The presence of outliers is expected; it is impossible to protect a priori against unrepresentative samplings, and with M samples in each of the $\mathcal{M}$ Monte Carlo instances the presence of unrepresentative sets $\{J_{T,hp}\}$ becomes likely when M is small. These unrepresentative samplings and the resulting outlier fits can generally be easily identified and, in practice, handled by user intervention. In the study in this section, any fit exhibiting q deviating by more than 0.5 or r deviating by more than 0.125 from the expected values or $J_{T,hp}$ deviating the prior guess by more than 2.0 are omitted. This constitutes 16 of the 1700 simulations in the forthcoming study.

We begin by evaluating how well the method estimates J_∞ with posterior estimates $(J_\infty)_\text{MAP}$. Figure 2.12 shows the convergence of the posterior output estimate $(J_\infty)_\text{MAP}$ to the reference value J_ref found in (1.23). From this plot, we can see that the estimator does a good job of estimating this quantity from $\{J_{T,hp}\}$, with below one

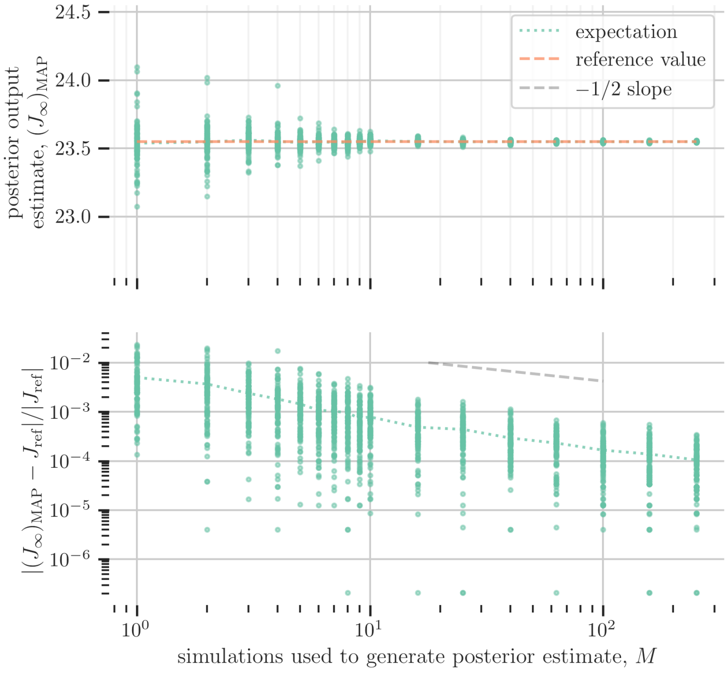


Figure 2.12: Monte Carlo results for output estimate $(J_\infty)_\text{MAP}$. Reference value J_ref and expected statistical convergence rate $M^{-1/2}$ are shown.

percent error for simulations with $M \gtrsim 10$. Error convergence seems to achieve the $M^{-1/2}$ rate expected in the size of the data for the Bayesian method asymptotically.

Another key result for the Bayesian method is to accurately approximate Δt_{opt}, such that the $(M+1)$-th simulation might be optimally discretized, as well as e_{opt}, which characterizes the error achievable at the given cost budget. In Figure 2.13, the estimates based on the length-M $\{J_{T,hp}\}$ for the optimizer are shown for each of the $\mathcal{M}$ Monte Carlo instances. This result demonstrates that the Bayesian small sample method, even with as few as ten simulations with randomized Δt can fairly accurately estimate the error and find the optimizer with small M. As a final measure of this ability, we can take the $\Delta t_{\text{opt,MAP}}$ using each θ^*_{MAP} in the Monte Carlo simulation and compare the reference error $e_{\text{model,ref}}$ against the optimal error from the reference model $e_{\text{opt,ref}}$. A plot of this excess error factor with probability density estimates is given in Figure 2.14. This result shows that the Bayesian error model, using only randomized Δt selection, can reliably identify near-optimal choices of Δt using a small number of simulations.

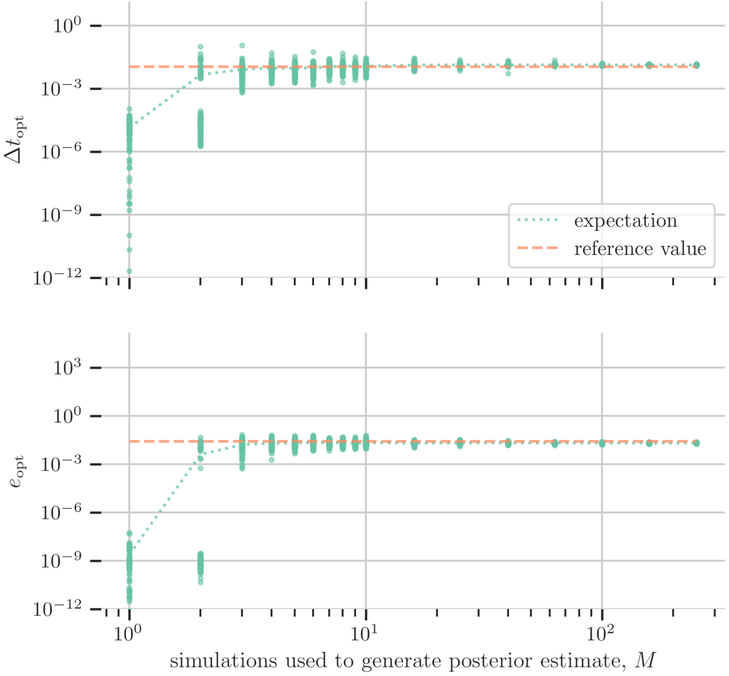


Figure 2.13: Monte Carlo results for optimizer and optimum of asymptotic model at $N_s = 1 \times 10^5$.

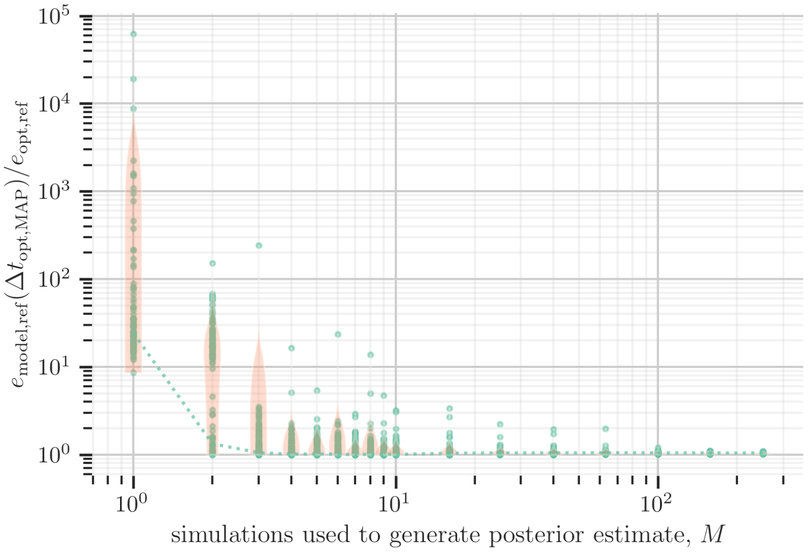


Figure 2.14: Monte Carlo results for excess error factor of asymptotic model at $N_s = 1 \times 10^5$. Orange regions give density estimates.

2.3 *Conclusions and future work*

In this chapter, we have shown that models to describe the error in discrete approximations $J_{T,hp}$ of an infinite-time mean output quantity J_∞ from a chaotic, ergodic system can be described in a way that is compatible with a Bayesian formulation. Using this Bayesian method, we show that it is possible to meaningfully estimate the error behavior, even with a small number of samples.

Having shown here that it is possible to generate approximate error models using a small-sample approach, now want to show in Chapter 3 that such an approach can be generalized to the case of ergodic, chaotic PDEs, where discretization effects from both temporal and spatial discretizations are present. From there, the keystone of the work will be to leverage the small sample approximation capability developed here while simulating a dynamical system. In general, we want to expend a computational budget to arrive at the best possible estimate given that budget of a quantity of interest like J_∞. Leveraging small-sample estimation capability to do so will be the concern of Chapter 4.

3

Error behavior and identification for ergodic chaotic partial differential equations

You can watch the tape,
You can try to hit your spots,
But don't do it for anything but
the love of movement and location,
Or the battle is lost.

—Punch Brothers, "Movement & Location"
Who's Feeling Young Now?

So far we have demonstrated that the initial transient behavior and the behavior on the attractor can be quantified for a chaotic, ergodic ODE. In this chapter, we will attempt to extend the error modeling and identification framework in Chapters 1 and 2 for chaotic partial differential equation (PDE) problems.

For PDEs, solution requires the discretization of both space and time. Many options for discretization exist, including prominently finite difference, finite volume, and finite element methods. All three remain common in industrial use, with benefits and costs for each. In this work, we will concentrate on the finite element method for spatial discretization combined with an implicit time integrator that uses a method of lines approach.

While we opt to use a method of lines temporal discretization with FEM spatial discretization here, the approach in this chapter can readily be extended to other discretization schemes, including space-time discretizations.

The use of provably stable FEM methods and implicit timestepping removes stability requirements that are common for other choices of schemes. Using the finite element approach on grids of elements with some characteristic size Δx, coupled with stable implicit time-marching methods that march timesteps separated by some Δt, allows choices about Δx and Δt to be primarily based on error considerations. In this chapter we will develop an error model to understand and control these errors in the setting of chaotic systems.

3.1 Error modeling

In order to efficiently estimate quantities of interest from spatio-temporally chaotic dynamical systems, we want to be able to control the various forms of error that enter into the result of a given simulation. We break the error induced by simulation into two separate categories. The first category, as previously, are the statistical errors, composed of spin-up error and sampling error. The second category, discretization errors, now includes contributions from the spatial discretization as well as the time-stepping/temporal advancement scheme. In this section, we will attempt to quantify each of these for PDEs.

Cost & error on the attractor

Consider a spatio-temporal dynamical system given by a PDE:

$$\frac{\partial \mathbf{u}(\vec{x},t)}{\partial t} = \mathbf{f}(\mathbf{u};\vec{x},t), \tag{3.1}$$

with an initial condition given by $\mathbf{u}(\vec{x},0) = \mathbf{u}_{\mathrm{IC}}(\vec{x})$, where $\mathbf{u}$ is the state of the system in $\mathbb{R}^n$, and $\mathbf{f}$ is a nonlinear spatial differential operator, $\vec{x} \in \mathbb{R}^d$, and $t \in \mathbb{R}$; d gives the number of spatial dimensions in the problem and n gives the number of states. In this work, we will first approximate the spatial system by an appropriate spatial discretization method, such that we formulate a problem of the form:

$$\frac{\mathrm{d}\mathbf{u}_{h_x p_x}(t)}{\mathrm{d}t} = \mathbf{f}_{h_x p_x}(\mathbf{u}_{h_x p_x};t), \tag{3.2}$$

where $\mathbf{u}_{h_x p_x}$ represents the solution of ordinary differential equation (ODE) system with $N_{\mathrm{DOF},x}$ degrees of freedom that results from the discrete spatial system characterized by a discretization scale $h_x = \Delta x$ and approximating order p_x, which we will define more explicitly later. Once we have this semi-discrete ODE form, we can then apply a method of lines discretization in time which can be characterized by a temporal scale $h_t = \Delta t$ and approximation order p_t.

As with ODE problems, we are frequently interested in estimating the mean value of output quantities of interest:

$$J_\infty = \lim_{T\to\infty} \frac{1}{T} \int_{t_0}^{t_0+T} g(\mathbf{u}(\vec{x},t))\, \mathrm{d}t, \tag{3.3}$$

where g is an instantaneous output functional of interest. Analogously, we approximate this value by a discrete approximation:

$$J_{T,hp} = \frac{1}{T_s} \mathbf{I}_{t_0}^{t_0+T_s} g_{hp}(\mathbf{u}_{hp}(\vec{x},t))\, \mathrm{d}t, \tag{3.4}$$

where $h = (h_x, h_t)$ and $p = (p_x, p_t)$ are multi-indices, and $\mathbf{u}_{hp}(\vec{x},t)$ represents the discrete solution function, which in general can be interpolated onto $(\vec{x},t)$ but in practice will be evaluated at discrete points $\vec{x}$ and t in space and time. As in Chapter 1, the integral is approximated discretely, and we assume that the effect of numerical integration of the output quantity is designed to be dominated by the discrete error effects.

Consider the case where t_0 is very large, and $\mathbf{u}_0$ has effectively converged onto the attractor. In this case, transient errors will be negligible and the remaining errors will be due to sampling and the spatial and temporal discretizations. We define the error induced by such a discrete estimate as $e_{T,hp}$:

$$e_{T,hp} \equiv J_{T,hp} - J_\infty. \tag{3.5}$$

Using an intermediate value

$$J_T = \frac{1}{T_s} \int_{t_0}^{t_0+T_s} g(\mathbf{u}(\vec{x},t))\, \mathrm{d}t, \tag{3.6}$$

we can write, equivalently:

$$e_{T,hp} \equiv \underbrace{(J_{T,hp} - J_T)}_{e_{hp}} + \underbrace{(J_T - J_\infty)}_{e_T}. \tag{3.7}$$

We can rewrite this term, breaking the discretization error into the sum of two components:

$$e_{T,hp} = \underbrace{e_{hp,x} + e_{hp,t}}_{e_{hp}} + e_T, \tag{3.8}$$

where $e_{hp,x}$ gives the spatial discretization error, $e_{hp,t}$ gives the temporal discretization error, and e_T gives the sampling error. Taking the absolute value, then using the triangle inequality and expectation function, we can write:

$$\mathbb{E}[|e_{T,hp}|] \leq \mathbb{E}[|e_{hp,x}|] + \mathbb{E}[|e_{hp,t}|] + \mathbb{E}[|e_T|], \tag{3.9}$$

where expectations are taken on the stationary distribution of the attractor. Following Chapter 1, we can use the central limit theorem (CLT) assuming satisfactorily strong mixing properties are present in the state of the dynamical system[1] such that (1.16) applies:

$$\mathbb{E}[|e_T|] \approx \frac{A_0}{\sqrt{M_{\text{ens}}}} T_s^{-r}. \tag{1.16}$$

Following Chapter 1, we can treat the temporal discretization error as in (1.19), assuming a form:

$$\mathbb{E}[|e_{hp,t}|] \approx B_{p_t} \Delta t^{q_t}, \tag{3.10}$$

with a positive parameter B_{p_t}. This leaves the new spatial discretization term, for which we will assume a form:

$$\mathbb{E}[|e_{hp,x}|] \approx C_{p_x} \Delta x^{q_x}, \tag{3.11}$$

with C_{p_x} a leading constant positive coefficient and q_x the relevant convergence order of the spatial discretization scheme.

Taking these three component models together, such that

$$\mathbb{E}[|e_{T,hp}|] \leq e_{\text{model}}, \tag{3.12}$$

we can write:

$$\begin{aligned} e_{\text{model}} &= e_{\text{model}}(\Delta x, \Delta t, T_s;\, M_{\text{ens}}, \theta) \\ &= C_{p_x} \Delta x^{q_x} + B_{p_t} \Delta t^{q_t} + \frac{1}{\sqrt{M_{\text{ens}}}} A_0 T_s^{-r}, \end{aligned} \tag{3.13}$$

where

$$\theta = \{q_x, q_t, r, C_{p_x}, B_{p_t}, A_0\} \tag{3.14}$$

gives the vector of relevant parameters.

Consider a budget $\mathcal{E}_s = N_{\text{elem}} N_s$ of "total elements computed" in a time-stepping scheme for sampling (i.e. starting on the attractor). Using $\mathcal{E}_s$, we can constrain T_s given Δx and Δt, since $N_{\text{elem}} = L/\Delta x$ and $N_s = T_s/\Delta t$

$$T_s = \frac{\mathcal{E}_s}{L} \Delta x \Delta t. \tag{3.15}$$

This allows us to insert (3.15) into (3.13). The result is minimized at:

$$\begin{aligned} \Delta x_{\text{opt}} = {} & q_x^{-\frac{q_t + r}{q_x q_t + q_x r + q_t r}} q_t^{\frac{r}{q_x q_t + q_x r + q_t r}} r^{\frac{q_t}{q_x q_t + q_x r + q_t r}} \\ & C_{p_x}^{-\frac{q_t + r}{q_x q_t + q_x r + q_t r}} B_{p_t}^{\frac{r}{q_x q_t + q_x r + q_t r}} A_0^{\frac{q_t}{q_x q_t + q_x r + q_t r}} \\ & M_{\text{ens}}^{-\frac{q_t}{2(q_x q_t + q_x r + q_t r)}} (\mathcal{E}_s / L)^{-\frac{q_t r}{q_x q_t + q_x r + q_t r}}, \end{aligned} \tag{3.16}$$

[1] Denker, 1989; and Bradley, 2005

$$\begin{aligned}\Delta t_{\text{opt}} = {} & q_x^{\frac{r}{q_xq_t+q_xr+q_tr}}\, q_t^{-\frac{q_x+r}{q_xq_t+q_xr+q_tr}}\, r^{\frac{q_x}{q_xq_t+q_xr+q_tr}} \\ & C_{p_x}^{\frac{r}{q_xq_t+q_xr+q_tr}}\, B_{p_t}^{-\frac{q_x+r}{q_xq_t+q_xr+q_tr}}\, A_0^{\frac{q_x}{q_xq_t+q_xr+q_tr}} \\ & M_{\text{ens}}^{-\frac{q_x}{2(q_xq_t+q_xr+q_tr)}}\, (\mathcal{E}_s/L)^{-\frac{q_xr}{q_xq_t+q_xr+q_tr}}\,,\end{aligned} \tag{3.17}$$

and

$$\begin{aligned}T_{s,\text{opt}} = {} & q_x^{-\frac{q_t}{q_xq_t+q_xr+q_tr}}\, q_t^{-\frac{q_x}{q_xq_t+q_xr+q_tr}}\, r^{\frac{q_x+q_t}{q_xq_t+q_xr+q_tr}} \\ & C_{p_x}^{-\frac{q_t}{q_xq_t+q_xr+q_tr}}\, B_{p_t}^{-\frac{q_x}{q_xq_t+q_xr+q_tr}}\, A_0^{\frac{q_x+q_t}{q_xq_t+q_xr+q_tr}} \\ & M_{\text{ens}}^{-\frac{q_x+q_t}{2(q_xq_t+q_xr+q_tr)}}\, (\mathcal{E}_s/L)^{\frac{q_xq_t}{q_xq_t+q_xr+q_tr}}\,,\end{aligned} \tag{3.18}$$

at which the error is given by:

$$\begin{aligned}e_{\text{opt}} = {} & (q_xq_t + q_xr + q_tr)\, q_x^{-\frac{q_xq_t+q_xr}{q_xq_t+q_xr+q_tr}}\, q_t^{-\frac{q_xq_t+q_tr}{q_xq_t+q_xr+q_tr}}\, r^{-\frac{q_xr+q_tr}{q_xq_t+q_xr+q_tr}} \\ & C_{p_x}^{\frac{q_tr}{q_xq_t+q_xr+q_tr}}\, B_{p_t}^{\frac{q_xr}{q_xq_t+q_xr+q_tr}}\, A_0^{\frac{q_xq_t}{q_xq_t+q_xr+q_tr}} \\ & M_{\text{ens}}^{-\frac{q_xq_t}{2(q_xq_t+q_xr+q_tr)}}\, (\mathcal{E}_s/L)^{-\frac{q_xq_tr}{q_xq_t+q_xr+q_tr}}\,,\end{aligned} \tag{3.19}$$

Thus we arrive at the best-possible error in expectation given a sampling budget $\mathcal{E}_s$.

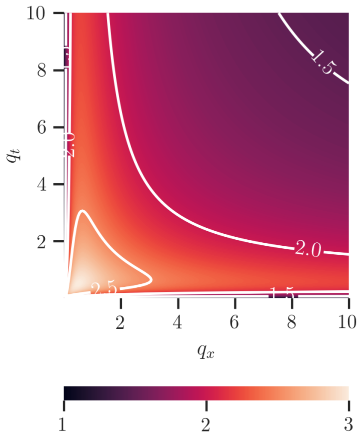


Figure 3.1: Consolidated product of terms in (3.19) exclusively composed of q_x, q_t, and r, assuming $r = 1/2$.

It is worth noting that the constant terms in the first line in (3.19) are equivalent to:

$$\left(\left(\frac{q_x}{r}\right)^{q_t}\left(\frac{q_t}{r}\right)^{q_x}\right)^{\frac{r}{q_xq_t+q_xr+q_tr}} + \left(\left(\frac{q_x}{q_t}\right)^{r}\left(\frac{r}{q_t}\right)^{q_x}\right)^{\frac{q_t}{q_xq_t+q_xr+q_tr}} + \left(\left(\frac{q_t}{q_x}\right)^{r}\left(\frac{r}{q_x}\right)^{q_t}\right)^{\frac{q_x}{q_xq_t+q_xr+q_tr}},$$

which is slightly more insightful; this term can be shown to be bounded between 1 and q for $r = 1/2$, and $q_x = q_t = q$. Values of this term can be seen in Figure 3.1.

Spin-up cost & error models

We now attempt to understand total cost and error of such a simulation including the pre-sampling cost and the spin-up error. In Section 1.5, we developed a model for the spin-up error for an ODE, which we will now use for the spatially discretized system given in (3.2). Specifically, we assume that the output behavior can be described by:

$$g(\mathbf{u}(\vec{x},t)) - g(\mathbf{u}^{\mathcal{A}}(\vec{x},t)) \approx A_\lambda \exp\left(-\frac{t}{T_\lambda}\right), \tag{3.20}$$

Here, A_λ represents the $t = 0$ displacement of the mean value of $g(\mathbf{u})$ from $g(\mathbf{u}^{\mathcal{A}})$, and T_λ represents the characteristic decay time of the deviation in g.

We can now integrate (1.52) from t_0 to $t_0 + T_s$ to get e_λ:

$$e_\lambda \equiv \frac{A_\lambda T_\lambda}{T_s} \exp\left(-\frac{t_0}{T_\lambda} \right). \tag{3.21}$$

A subtlety here is that, for an PDE system reduced to a large-degree-of-freedom ODE system, the likelihood of having complex, multi-modal decay is high, and this method might more appropriately understood to describe the decay of the dominant CLV mode. The true nature of this decay is generally obscured by the output functional g, however, and for simplicity and ease of interpretation we assume one dominant modal behavior.

Thus, the total errors can be described by

$$e_{T,hp} = e_{hp,x} + e_{hp,t} + e_\lambda + e_T. \tag{3.22}$$

Taking the absolute value, using the triangle inequality, and leveraging linearity of expectations, we find:

$$\mathbb{E}[|e_{T,hp}|] \le \mathbb{E}[|e_{hp,x}|] + \mathbb{E}[|e_{hp,t}|] + \mathbb{E}[|e_\lambda|] + \mathbb{E}[|e_T|], \tag{3.23}$$

where the appropriate expectations are taken on the initial conditions or on the attractor for each term. Using the models given in (1.16), (3.10), (3.11), and (1.52) we arrive at an error model that includes the effect of spin-up:

$$\begin{aligned} e_{\text{model}} &= e_{\text{model}}(\Delta x, \Delta t, t_0, T_s;\, M_{\text{ens}}, \theta, \psi) \\ &= C_{q_x} \Delta x^{q_x} + B_{q_t} \Delta t^{q_t} + \frac{A_0}{\sqrt{M_{\text{ens}}}} T_s^{-r} + |A_\lambda| \frac{T_\lambda}{T_s} \exp\left(-\frac{t_0}{T_\lambda} \right). \end{aligned} \tag{3.24}$$

Consider now total cost, including the spin-up time t_0, in terms of the total number of elements on which a solution is computed over all the timesteps. In this case, a total computed element budget will consist of spin-up as well as sampling costs:

$$\mathcal{E}_t = N_{\text{elem}} N_t = N_{\text{elem}} (N_0 + N_s). \tag{3.25}$$

This is a model for the error that is effective in particular for discontinuous Galerkin discretizations, but may be less so for other types of discretizations.

We can compute the sampling time:

$$T_s = \frac{\mathcal{E}_t}{L} \Delta x \Delta t - t_0, \tag{3.26}$$

and inserting into (3.24) gives:

$$\begin{aligned} e_{\text{model}} &= e_{\text{model}}(\Delta x, \Delta t, t_0;\, \mathcal{E}, M_{\text{ens}}, \theta, \psi) \\ &= C_{q_x} \Delta x^{q_x} + B_{q_t} \Delta t^{q_t} + \frac{A_0}{\sqrt{M_{\text{ens}}}} \left(\frac{\mathcal{E}_t}{L} \Delta x \Delta t - t_0 \right)^{-r} \\ &\quad + |A_\lambda| T_\lambda \left(\frac{\mathcal{E}_t}{L} \Delta x \Delta t - t_0 \right)^{-1} \exp\left(-\frac{t_0}{T_\lambda} \right), \end{aligned} \tag{3.27}$$

which is a model for the error as a function of Δx, Δt, and t_0: the three free choices in a given simulation for a given total elemental budget $\mathcal{E}_t$.

3.2 *Output likelihood modeling for small-sample estimation*

In practice, we will not have a priori access to the set of sampling error model parameters, θ, nor the transient model ψ, and least of all the true output J_∞, which we take as the primary aim of simulation. In (3.20), we described the transient behavior as a property of the instantaneous output quantity of interest. We will show in Section 3.4 a fitting procedure can identify ψ using the output trace of a given simulation. In this section, we will develop a modeling approach that can allow us to make posterior estimates of θ and J_∞ at low cost, allowing for the complete estimation of the cost-error relationship on reasonable computational budgets.

In order to identify θ, we revisit the assumptions used to generate the sampling error model. We start with the definitions in (3.8). We can combine (3.5) and (3.8) to find:

$$J_{T,hp} = J_\infty + e_{hp,x} + e_{hp,t} + e_T \tag{3.28}$$

We follow Chapter 2, incorporating spatial discretization effects. This results in discretization error models:

$$\begin{align} e_{hp,x} &\approx C^*_{p_x} \Delta x^{q_x} \tag{3.29} \\ e_{hp,t} &\approx B^*_{p_t} \Delta t^{q_t}, \tag{3.30} \end{align}$$

with $C^*_{p_x} \in \mathbb{R}$ and $B^*_{p_t} \in \mathbb{R}$ (as we are modeling $e_{hp,x}$ and $e_{hp,t}$ as opposed to $|e_{hp,x}|$ and $|e_{hp,t}|$ and the constants are no longer required to be positive). Taking (3.29), (3.30), and (2.7) and inserting into (3.28), we arrive at a random variable form for the output of a given simulation:

$$J_{T,hp} = \mathcal{N}\left(J_\infty + C^*_{p_x} \Delta x^{q_x} + B^*_{p_t} \Delta t^{q_t}, \left(A^*_0 T_s^{-r}\right)^2\right). \tag{3.31}$$

This allows us to describe the error of interest in expectation as well:

$$\begin{align} \mathbb{E}[|e_{T,hp}|] &= \mathbb{E}[|J_{T,hp} - J_\infty|] \\ &= \mathbb{E}\left[\left|\mathcal{N}\left(C^*_{p_x} \Delta x^{q_x} + B^*_{p_t} \Delta t^{q_t}, \left(A^*_0 T_s^{-r}\right)^2\right)\right|\right] \\ &\leq |C^*_{p_x} \Delta x^{q_x} + B^*_{p_t} \Delta t^{q_t}| + \mathbb{E}\left[\left|\mathcal{N}\left(0, \left(A^*_0 T_s^{-r}\right)^2\right)\right|\right] \tag{3.32} \\ &\leq |C^*_{p_x}| \, \Delta x^{q_x} + |B^*_{p_t}| \, \Delta t^{q_t} + \sqrt{\frac{2}{\pi}} A^*_0 T_s^{-r}, \end{align}$$

where we have applied the assumption that the discretization errors have a deterministic effect alongside the triangle inequality and the expectation of a half-normal variable. This allows the insight that our likelihood-based model arrives at a form that is equivalent to (3.13), under the transformation:

$$C_{p_x} = |C^*_{p_x}| \qquad B_{p_t} = |B^*_{p_t}| \qquad A_0 = \sqrt{\frac{2}{\pi}} A^*_0. \tag{3.33}$$

Thus, if we can make an estimate to the parameters

$$\theta^* = \left\{ C^*_{p_x}, q_x, B^*_{p_t}, q_t, A^*_0, r, J^*_\infty \right\}, \tag{3.34}$$

we can translate this result into an estimate of θ and (3.13) via the parameter mapping in (3.33).

We can now use (3.31) to write the probability of any given series of output results:

$$p\left(\left\{ J^{(i)}_{T,hp} \right\} \middle| J_\infty, C^*_{p_x}, q_x, B^*_{p_t}, q_t, A^*_0, r \right) = \prod_{i=1}^{N_{\text{samp}}} \mathcal{N}\left(J^{(i)}_{T,hp}; J_\infty + C^*_{p_x} \Delta x^{q_x} + B^*_{p_t} \Delta t^{q_t}, \left(A^*_0 T_s^{-r} \right)^2 \right), \tag{3.35}$$

since any two output averages are independently generated. In this work, we will assume that p_x, p_t, and r satisfy the theoretical rates. Thus, using Bayes rule, we can write:

$$p\left(J_\infty, C^*_{p_x}, B^*_{p_t}, A^*_0 \middle| \left\{ J^{(i)}_{T,hp} \right\} \right) \propto p\left(\left\{ J^{(i)}_{T,hp} \right\} \middle| J_\infty, C^*_{p_x}, B^*_{p_t}, A^*_0 \right) p\left(J_\infty, C^*_{p_x}, B^*_{p_t}, A^*_0 \right). \tag{3.36}$$

The implication of this statement is that, with a set of results $\{J^{(i)}_{T,hp}\}$ and a prior model $p\left(J_\infty, C^*_{p_x}, B^*_{p_t}, A^*_0 \right)$, we can create an estimate for the model parameters J_∞, $C^*_{p_x}$, $B^*_{p_t}$, and A^*_0.

We now generate a prior model. For J_∞, we generally have an intuition for a quantity of interest; while we don't know *exactly* what we expect J_∞ to be, we can guess a plausible range. In this case, we can make a guess that the average value of $g = \int_\Omega u^2(\vec{x}, t)\, \mathrm{d}\Omega$ is about 100 ± 50. We can codify this by a normal distribution:

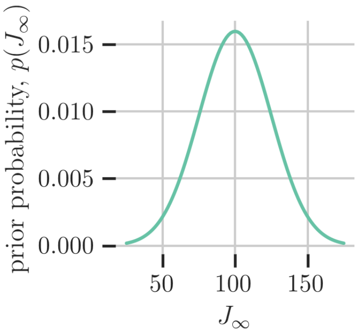


Figure 3.2: Prior distribution for J_∞.

In the case of this work, this guess is generated from knowledge of prerequisite results generated in the process of the research. In general, these types of estimates can be made by literature study (e.g. Blonigan and Wang [2014] for the modified KSE used here), extrapolation of empirical results, or low-fidelity models.

$$p\left(J_\infty \right) = \mathcal{N}\left(\mu_J, \sigma_J^2 \right) \tag{3.37}$$

with $\mu_J = 100$ and $\sigma_J = 25$. This distribution can be seen in Figure 3.2.

On the other hand, we want to assume that variables $C^*_{p_x}$, $B^*_{p_t}$, and A^*_0 are unknown, at least up to a order of magnitude. To achieve this, we use a large variance, zero-mean normal distribution for $C^*_{p_x}$ and $B^*_{p_t}$:

$$p\left(C_{p_x} \right) = \mathcal{N}\left(0, \sigma_C^2 \right) \tag{3.38}$$

and

$$p\left(B_{p_t}\right) = \mathcal{N}\left(0, \sigma_B^2\right), \tag{3.39}$$

with $\sigma_C = \sigma_B = 10^5$. This distribution can be seen in Figure 3.3, and we note that there is a non-negligible probability associated with magnitudes of C_{p_x} and B_{p_t} up to 10^6.

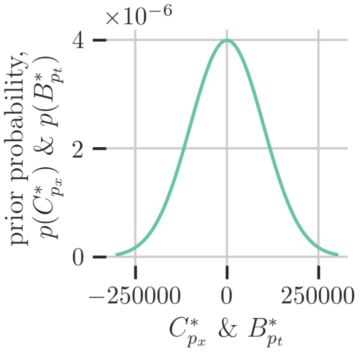


Figure 3.3: Prior distribution for $C^*_{p_x}$ and $B^*_{p_t}$.

For A_0^*, which is non-negative, we use a halfnormal distribution:

$$p\left(A_0^*\right) = \mathcal{H}\left(\sigma_A^2\right) \tag{3.40}$$

with $\sigma_A = 100$. This distribution can be seen visualized in Figure 3.4.

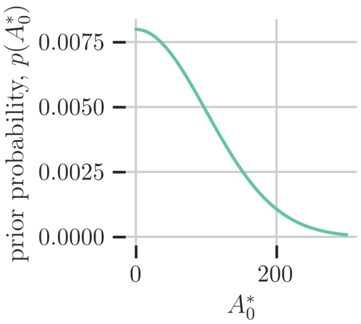


Figure 3.4: Prior distribution for A_0^*.

These values are chosen to have orders of magnitude significantly larger- by at least two orders of magnitude- of the values we found at high cost in Table 3.5. Taking these four variables as independent, results in a prior model:

$$p\left(J_\infty, C^*_{p_x}, B^*_{p_t}, A^*_0\right) = p\left(J_\infty\right) p\left(C^*_{p_x}\right) p\left(B^*_{p_t}\right) p\left(A^*_0\right). \tag{3.41}$$

Thus, (3.36) can be fully specified, and Bayesian posterior estimates can be made of the model parameters J_∞, $C^*_{p_x}$, $B^*_{p_t}$, and A^*_0.

3.3 *Kuramoto-Sivashinsky equation*

In order to evaluate the models proposed in this work, we will solve the Kuramoto-Sivashinsky equation (KSE) with convection, given by:

$$\begin{aligned} \frac{\partial u}{\partial t} + (\boldsymbol{c} + \boldsymbol{\alpha} u) \cdot \nabla u + \beta \nabla^2 u + \gamma \nabla^4 u = 0 \qquad & \text{in } \Omega \\ u = \nabla u \cdot \hat{n} = 0 \qquad & \text{on } \partial\Omega \end{aligned} \tag{3.42}$$

with $\boldsymbol{c} = 1.6$, $\boldsymbol{\alpha} = 1$, $\beta = 1$, $\gamma = 1$. The boundary conditions in (3.42) are clamped-plate BCs and are necessary for realizing a chaotic solutions. In this work, we will used a one-dimensional ($d = 1$) domain Ω from $x = 0$ to $x = L$ with $L = 128$, for which the KSE is known to be chaotic with this set of parameters. The Kuramoto-Sivashinsky equation describes a variety of physical processes, with its first derivation for use in describing flame front propagation[2]. As a numerical model, the Kuramoto-Sivashinsky equation is notable as a simple problem which exhibits multi-scale spatiotemporal chaos, and for this reason is commonly used as a test problem for research in chaotic dynamical systems[3].

In Figures 3.5 and 3.6, we show two sample solutions of the Kuramoto-Sivashinsky equation. Figure 3.5 shows the behavior from $t = 0$ to

[2] Yoshiki Kuramoto. Diffusion-induced chaos in reaction systems. *Progress of Theoretical Physics Supplement*, 64: 346–367, 02 1978; and Gregory I. Sivashinsky. Nonlinear analysis of hydrodynamic instability in laminar flames– I. derivation of basic equations. *Acta Astronautica*, 4(11):1177–1206, 1977

[3] Blonigan and Wang, 2014; and Johan Larsson. Grid-adaptation for chaotic multi-scale simulations as a verification-driven inverse problem. In *AIAA Aerospace Sciences Meeting*, 2018

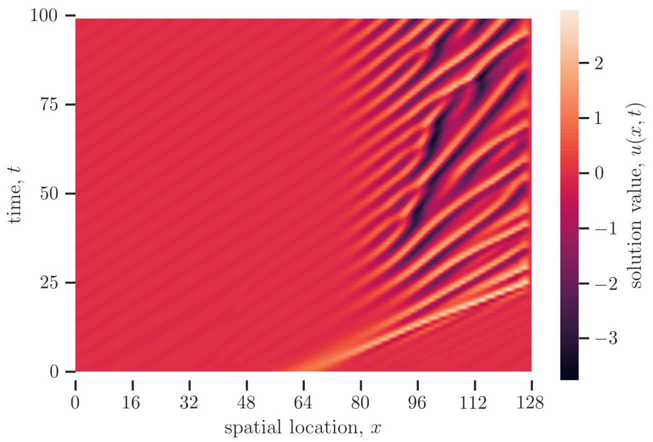


Figure 3.5: Solution of modified Kuramoto-Sivashinsky equation with $0 \leq t \leq 100$.

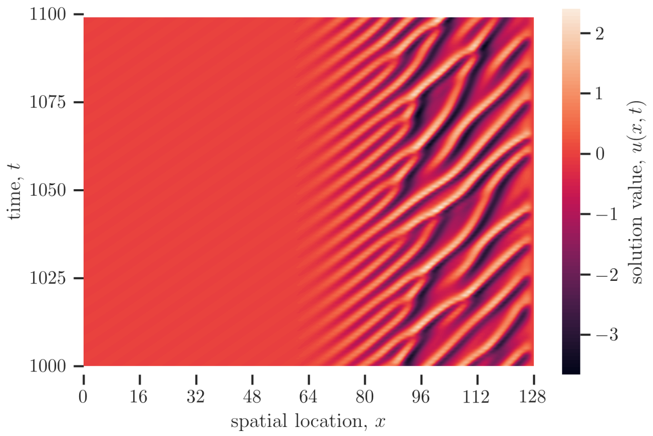


Figure 3.6: Solution of modified Kuramoto-Sivashinsky equation with $1000 \leq t \leq 1100$.

$t = 100$, as the KSE develops after being initialized at $t = 0$ with a Gaussian initial condition given by:

$$u_{\mathrm{IC}}(x) = \exp\left(\left(\frac{x - L/2}{L/32}\right)^2\right). \tag{3.43}$$

On the other hand, Figure 3.6 gives the behavior from $t = 1000$ to $t = 1100$, which is more characteristic of the statistically stationary behavior. The resulting solutions demonstrate spatial development of the linear instabilities as the solution is convected by $\boldsymbol{c}$ from left-to-right. These instabilities result in aperiodic and unpredictable coherent structures in the region with $x \gtrsim 64$ which result in consistent and stationary mean behavior over long times.

In Figure 3.7, the instantaneous energy output functional of the system,

$$g(\mathbf{u}(\vec{x},t)) = \int_\Omega u^2(\vec{x},t)\ \mathrm{d}\Omega, \tag{3.44}$$

is shown, with the sampling periods from Figure 3.5 and Figure 3.6

both shown. In this plot, the emergence of statistically stationary behavior can be seen, along with evident spin-up transient behavior, which are of critical interest for making accurate estimates of long-term averages.

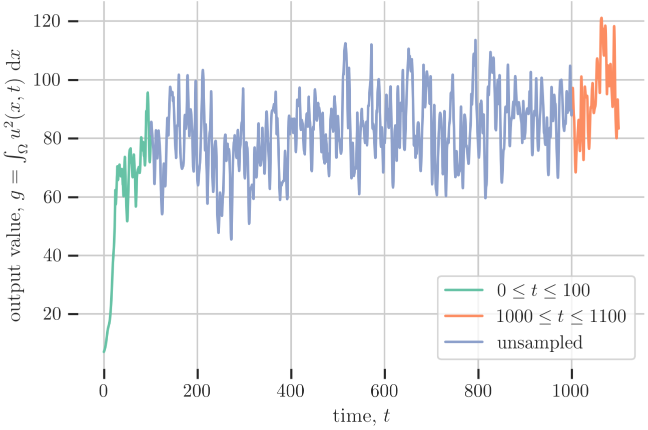


Figure 3.7: Output trace of modified Kuramoto-Sivashinsky equation.

Discretization method

In this work, we will solve the spatial behavior of the KSE using a new discontinuous Galerkin (DG) scheme that extends the second method of Bassi and Rebay[4] (DGBR2) to fourth-order operators solved with a Ciarlet-Raviart auxiliary-variable form[5]. We will refer to this scheme as a "DGBR4" discretization. The DGBR4 scheme gives a finite element solution of the spatial physics with order p_x polynomial solutions on the elements. The resulting semi-discrete differential-algebraic form is then discretized in time by a DIRK-DAE scheme that advances the state with a valid approximation of the auxiliary variable. Complete details on the scheme can be found in Appendix B.

The classic analysis for DG schemes[6] results in global error convergence rates for non-chaotic problems that scale as Δx^{p_x+1}. Accordingly, we will assume that the output functional g is bounded and thus that $q_x = p_x + 1$. The temporal discretization error convergence rate for the RK scheme is expected to be $q_t = p_t$.

Initial conditions for the forthcoming Kuramoto-Sivashinsky simulations in this work use a randomly-perturbed initial condition, in which each degree of freedom is chosen from a zero-mean random variable with a standard deviation given by ϵ, with $\epsilon = 10^{-6}$ everywhere.

[4] F. Bassi and S. Rebay. A high-order accurate discontinuous finite element method for the numerical solution of the compressible Navier-Stokes equations. *Journal of Computational Physics*, 131(2):267–279, 1997

[5] Philippe G. Ciarlet and Pierre-Arnaud Raviart. A mixed finite element method for the biharmonic equation. In Carl de Boor, editor, *Mathematical Aspects of Finite Elements in Partial Differential Equations*, pages 125–145. Academic Press, 1974

[6] Cockburn et al., 2000

In addition to global error convergence rates, superconvergent rates in output functionals have been shown to exist when PDEs are discretized in an adjoint consistent and fully variational manner [Pierce and Giles, 2000]. This type of discretization is outside the scope of this work, but this framework can be extended to investigate the effect of superconvergent methods for chaotic systems.

Computational cost

We now seek to quantify the computational cost of generating any given solution of the Kuramoto-Sivashinsky equation using the DGBR4/DIRK-DAE scheme. We start by noting that the spatial DGBR4 scheme for the 1D Kuramoto-Sivashinsky equation linearizes to a block triangular system. Thus, the spatial system can be solved with a block Thomas algorithm to handle inter-element interactions, with Gaussian elimination on element-wise blocks for the intra-element interactions. We assume that the cost of solving the spatial nonlinear system on average is well approximated by some constant times the cost of the inversion of the linearized system. Each nonlinear solve of the spatial system happens within the implicit Runge-Kutta scheme. For the DIRK-DAE schemes used in this work, each timestep will require p_t inversions of the spatial system to advance the primary state, as well as one additional final inversion that advances the auxiliary state. Combining these assumptions, the cost of the scheme is expected to scale with

We note that this advancement scheme is likely not the most efficient approach possible for this type of system.

$$\begin{aligned}\mathcal{C}_t &= \underbrace{\left[(p_x+1)^3 N_{\text{elem}}\right]}_{\text{cost of linear solve}} \underbrace{\left[(p_t+1)N_t\right]}_{\text{timestepping cost}} \\ &= \mathcal{C}_{p_x p_t} N_{\text{elem}} N_t = \mathcal{C}_{p_x p_t} \mathcal{E}_t,\end{aligned} \tag{3.45}$$

and thus the cost of this scheme will scale with the total number of elements computed, $\mathcal{E}_t = N_{\text{elem}} N_t$, where

$$\mathcal{C}_{p_x p_t} = (p_x+1)^3 (p_t+1) \tag{3.46}$$

is a discretization-dependent constant. Experiments validating this cost model can be found in Appendix C.

3.4 *Numerical experiments*

We will now perform a series of computations to show that the (3.13), (3.27), and (3.35) have explanatory value for simulations of the Kuramoto-Sivashinsky equation.

Reference estimate of J_∞

We start by making a reference approximation $J_{\text{ref}} \approx J_\infty$ that can be used in turn to generate approximations of the error. In order to estimate this value, we must assume a value for the spin-up time, t_0. We set the spin-up time to $t_0 = 1.71 \times 10^4$, then simulate for another $T_s = 29{,}965$ using $N_s = 1{,}004{,}214$ timesteps, such that $\Delta t \approx 2.98 \times 10^{-2}$. The same timestep is used for the spin-up time. The physical domain is discretized using $N_x = 1297$ equispaced elements over the domain Ω. These parameters are used with the $p_x = 2$ DG scheme and the 3^{rd}-order Runge-Kutta scheme. We generate an ensemble of $M_{\mathrm{ens}} = 82$ instances to estimate the true value.

The result of this simulation is an estimate of the output of interest, which in this work will everywhere be the energy of the KSE system given in (3.44). This gives an estimate of J_∞:

$$J_{\mathrm{ref}} = 118.44 \pm 0.05, \tag{3.47}$$

with the standard error given based on the ensemble estimate.

Error model reference values

We now simulate the KSE over a grid of choices of Δx, Δt, and T_s in order to identify θ and θ^*. In order to ensure fitting of the asymptotic behavior, we have observed approximate limits of the asymptotic convergence region, given in Table 3.1. Given these maximum values,

Δx_{max}	0.24
Δt_{max}	0.50
$T_{s,\mathrm{min}}$	10

(a) $p_x = 1$ DG, RK2

Δx_{max}	1.00
Δt_{max}	0.24
$T_{s,\mathrm{min}}$	10

(b) $p_x = 2$ DG, RK3

Table 3.1: Observed limits of asymptotic convergence.

we simulate on a grid of Δx and Δt values that span one order of magnitude, setting T_s at each $(\Delta x, \Delta t)$ pair under the constraint in (3.15) at a fixed $\mathcal{C}_s = 10^7$, where $\mathcal{C}_s$ is defined analogously to $\mathcal{C}_t$ over the N_s timesteps used for sampling (rather than the total number of timesteps N_t). All simulations are run with $t_0 = 12{,}000$; as Δx and Δt get smaller, the number of spin-up timesteps grows at a high rate. At each $(\Delta x, \Delta t)$ point, we run 100 simulations with the zero-mean randomly-perturbed initial condition. Using the estimates $J_{T,hp}$ and the reference estimate of J_{ref} from (3.47), we approximate $\mathbb{E}[|e_{T,hp}|]$ by the Monte Carlo method.

Next, we seek reference values of $C^*_{p_x}$, $B^*_{p_t}$, and A^*_0 (and, in turn, C_{p_x}, B_{p_t}, and A_0). In order to generate reference values for these, (3.35) can rearranged such that $e_{T,hp}$ is given by a normal variable with discretization-dependent mean

$$\mu_J(\Delta x, \Delta t) = C^*_{p_x} \Delta x^{q_x} + B^*_{p_t} \Delta t^{q_t}, \tag{3.48}$$

and a standard deviation that varies as $A^*_0 T_s^{-1/2}$ (in the limit as $r \to 1/2$). Since we know the standard deviation is $T_s^{-1/2}$ under the CLT up to a constant, we use a weighted nonlinear least squares fit to find $C^*_{p_x}$ and $B^*_{p_t}$ from (3.48). Next, we use the quantity

$$\frac{e_{T,hp} - \mu_J}{\sqrt{T_s}}$$

to identify A_0^*, since this quantity should be a zero-mean normal distribution with standard deviation A_0^*. On the top plots of Figures 3.8 and 3.9, $e_{T,hp}$ is shown with the respective discretization error behavior. The lower left plots show the corrected error quantity $(e_{T,hp} - \mu_J)$ and two standard deviations about zero, which show some pre-convergent[7] behavior under the CLT for $T_s < 50$, which we exclude from the fits for that reason. In the bottom left plot, we show the distribution of the non-dimensional error quantity for $T_s < 50$, overlaid with a Gaussian function having the same standard deviation.

$C^*_{p_x}$	480.418
q_x	2
$B^*_{p_t}$	−47.541
q_t	2
A_0^*	71.571
r	1/2

Table 3.2: Result of θ^* fit for discrete estimates of mean KSE energy output using $p_x = 1$ DG, RK2 discretization.

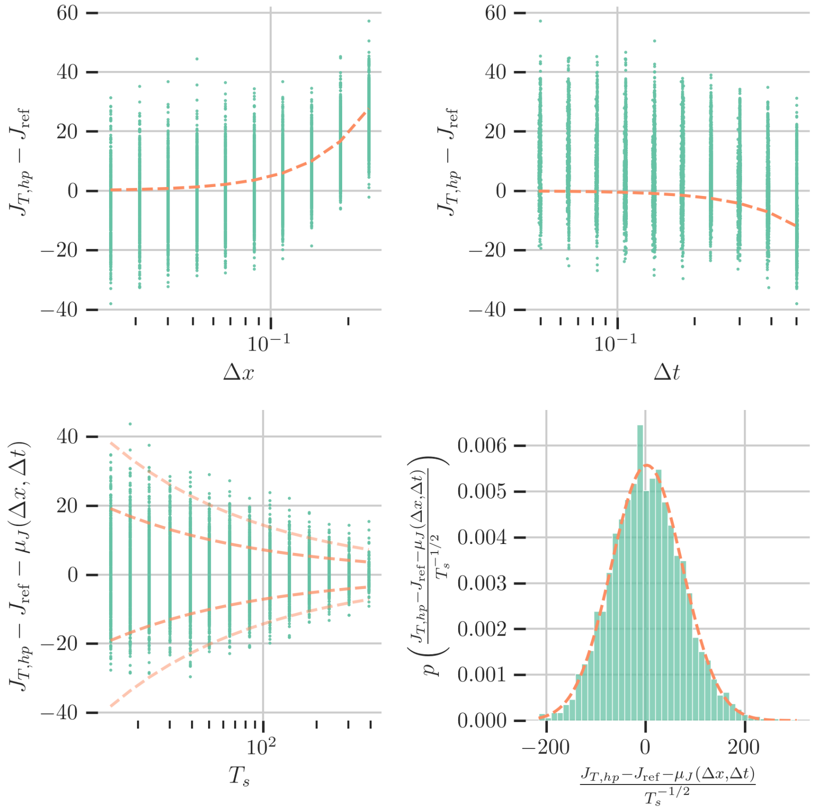


Figure 3.8: Error behavior and reference fits for likelihood model parameters for $p_x = 1$, RK2 simulation of KSE energy. Fits and histogram for $T_s < 50$.

Because we expect the variation to be a property of the physical system, independent of the discretization, we can also directly compare the non-dimensionalized error quantity between the two discretizations, allowing a combined estimate of A_0^* and the histogram in Figure 3.10.

$C^*_{p_x}$	15.640
q_x	3
$B^*_{p_t}$	−749.493
q_t	3
A_0^*	63.983
r	1/2

Table 3.3: Result of θ^* fit for discrete estimates of mean KSE energy output using $p_x = 2$ DG, RK3 discretization.

[7] It is very possible that the observed pre-convergent behavior in T_s is due to insufficient spin-up at $t_0 = 12{,}000$.

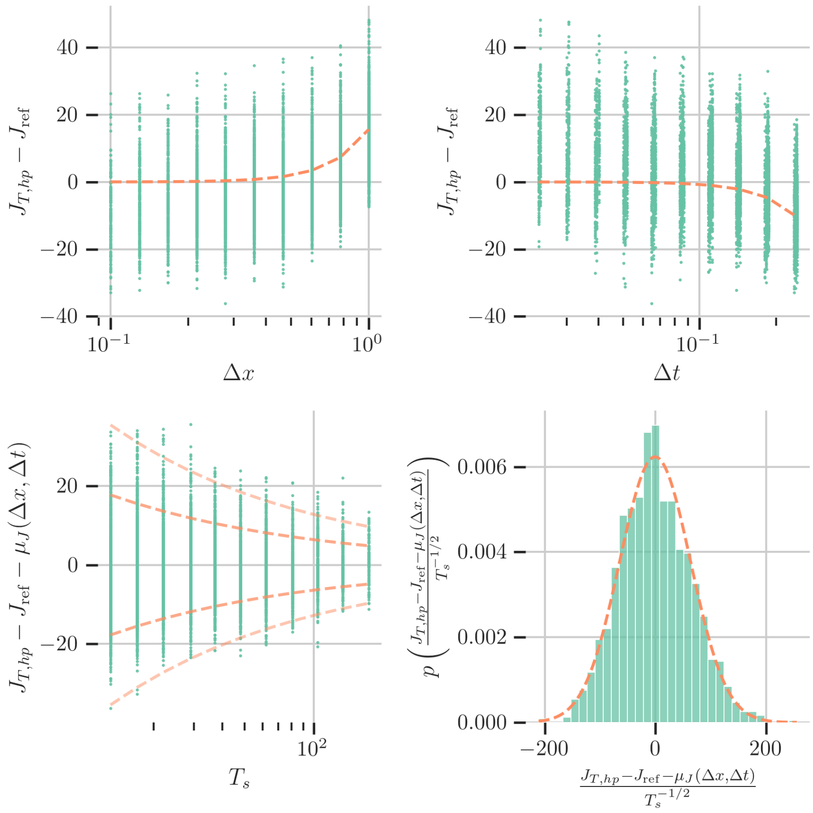


Figure 3.9: Error behavior and reference fits for likelihood model parameters for $p_x = 2$, RK3 simulation of KSE energy. Fits and histogram for $T_s < 50$.

A_0^*	69.342
r	1/2

Table 3.4: Combined statistical fit for discrete estimates of mean KSE energy output.

The resulting reference values for θ^* can be found in Tables 3.2, 3.3, and 3.4. In addition to θ^*, we can also use (3.33) to compute the implied values of θ, which we give in Table 3.5.

C_{p_x}	480.418
q_x	2
B_{p_t}	47.541
q_t	2
A_0	57.105
r	1/2

(a) $p_x = 1$ DG, RK2

C_{p_x}	15.640
q_x	3
B_{p_t}	749.493
q_t	3
A_0	51.051
r	1/2

(b) $p_x = 2$ DG, RK3

Table 3.5: Estimates of θ to characterize error model for discrete estimates of mean KSE energy output.

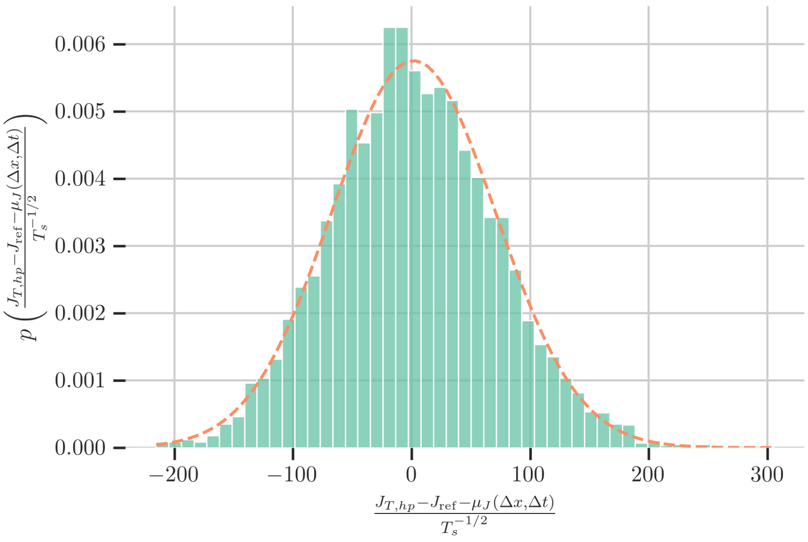


Figure 3.10: Reference fit for likelihood model parameter A_0^* from combined $p_x = 1$, RK2 and $p_x = 2$, RK3 discretizations of KSE energy. Fits and histogram for $T_s < 50$.

Spin-up model fitting

Having identified a model for the behavior of the error on the attractor, we now must characterize the spin-up transient in order to completely understand the relationship between cost and error. For a reference value of the underlying spin-up behavior, we will study the outputs of the reference cases used to generate (3.47). Our goal is to identify ψ for the transient behavior.

We follow the procedure outlined in Section 1.5, computing the likelihood of the output trace using (1.59) and using the priors in (1.60). We use different hyperparameters for the T_λ prior, specialized for the KSE traces we observe here:

$$(\alpha_T, \beta_T) \Longleftarrow (\mu_T = 1000.0, \sigma_T = 500.0)\,.$$

In Figure 3.11 we show a sample of the fit for $g(t) = \int_\Omega u^2(\vec{x}, t)\, \mathrm{d}\Omega$ with the $p_x = 2$, RK3 solution. The result of this approximation is an estimate

$$\psi^* = (T_\lambda, A_\lambda) \approx (1730, -48.5), \tag{3.49}$$

from which values for ψ can be derived. We can repeat this process across the ensemble used to generate (3.47), and in Figure 3.12, we show a histogram of the results for the entire ensemble used for the reference estimate. There results suggest that a conservative estimate for the transient decay model for the KSE energy output with the

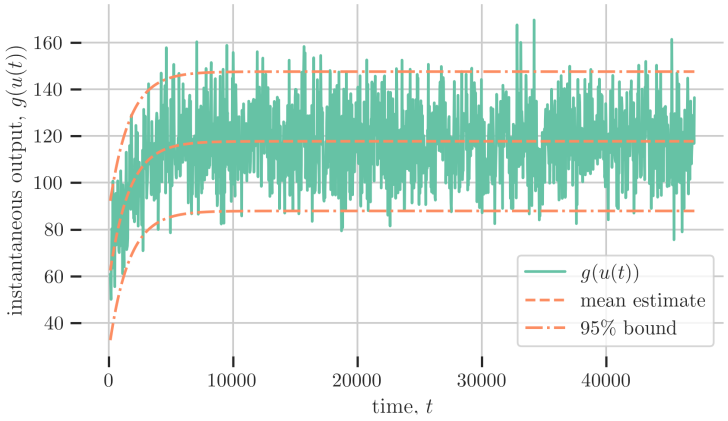
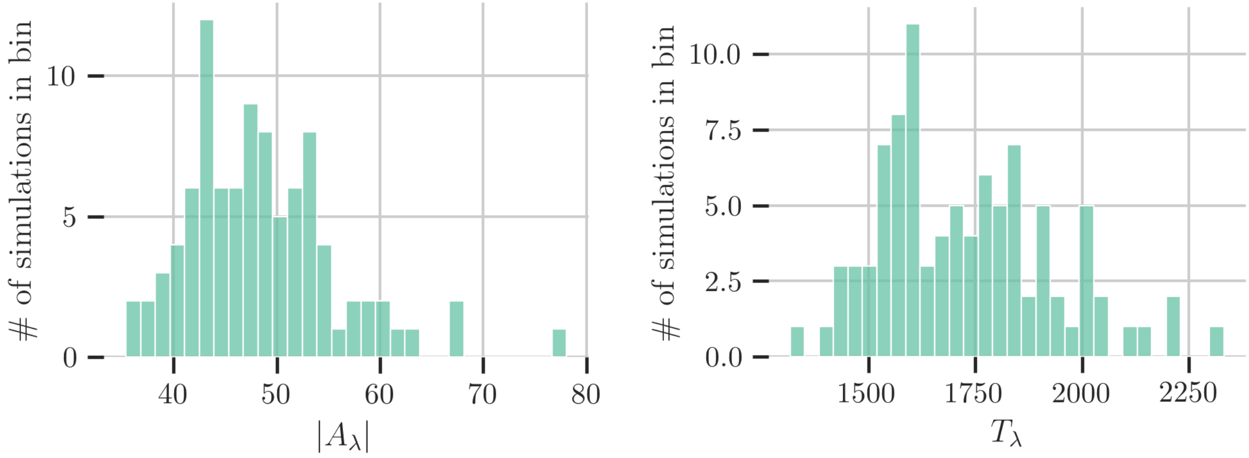


Figure 3.11: Sample trace of reference KSE energy output ($g(\mathbf{u}(t)) = \int_\Omega \mathbf{u}^2 \, \mathrm{d}\Omega$) with transient model fit. Results: $(J_\infty)_{\text{est}} = 117.72$, $\sigma_{g,\text{est}} = 14.900$, $T_{\lambda,\text{est}} = 1583.24$, $A_{\lambda,\text{est}} = -59.614$.



(a) $|A_\lambda|$

(b) T_λ

Figure 3.12: Histograms of transient model fits of ψ using reference ensemble.

fuzz initial condition:

$$\psi_{\text{ref}} = (T_\lambda, |A_\lambda|) \approx (2000, 60), \tag{3.50}$$

which we will use to characterize the spin-up transient for the purposes of error modeling.

Taken in sum, this section demonstrates that with the use of a Bayesian method we can make an estimate of the dominant exponential convergence behavior in any given ergodic trace of the Kuramoto-Sivashinsky system without strong a priori knowledge. This means, in turn, that we should be able to relate any choice of t_0 and T_s to an estimate of the error incurred by the transient convergence process using (1.52).

3.5 *Optimal simulation including spin-up*

Now that we have characterized the discretization, sampling, and transient errors associated with simulations of a given PDE output, we explore simulation optimality as a function of computational cost. Consider a simulation with a budget of $N_x N_t = 10^{12}$ total computed elements. We can begin by studying the effect of the choice of Δx and Δt in terms of the error under (3.27) with the estimates of ψ given in (3.50) and θ in Table 3.5. At each $(\Delta x, \Delta t)$, we compute the optimal spin-up time by mimimizing (3.27) under the realizability constraint $T > t_0$. In Figure 3.13, the error under the model is shown with an optimum value for the $p_x = 1$ DG/RK2 with $M_{\text{ens}} = 1$ and $M_{\text{ens}} = 64$. In Figure 3.14, an equivalent plot is shown for the $p_x = 2$ DG/RK3 case.

We observe that, as expected, the error models have clear and evident optima. Moving diagonally from top-right to bottom-left, we vary the total simulation time available under the budgets. We can see from the overlaid t_0 contours that the spin-up time is significant in the top-right region, where total simulation time is least constrained by the budget. As noted in Chapter 1, there are two modes for reduction of the spin-up errors: increasing T_s to reduce the spin-up error as T_s^{-1}, or increasing t_0 to decrease this error exponentially. Because the available $T = t_0 + T_s$ is set by Δx and Δt under the budget, when Δx and Δt are small, T must also be small. In this case, it becomes advantageous to use less t_0 in order to use the available T for sampling, which mutually controls the sampling error as well as the spin-up error.

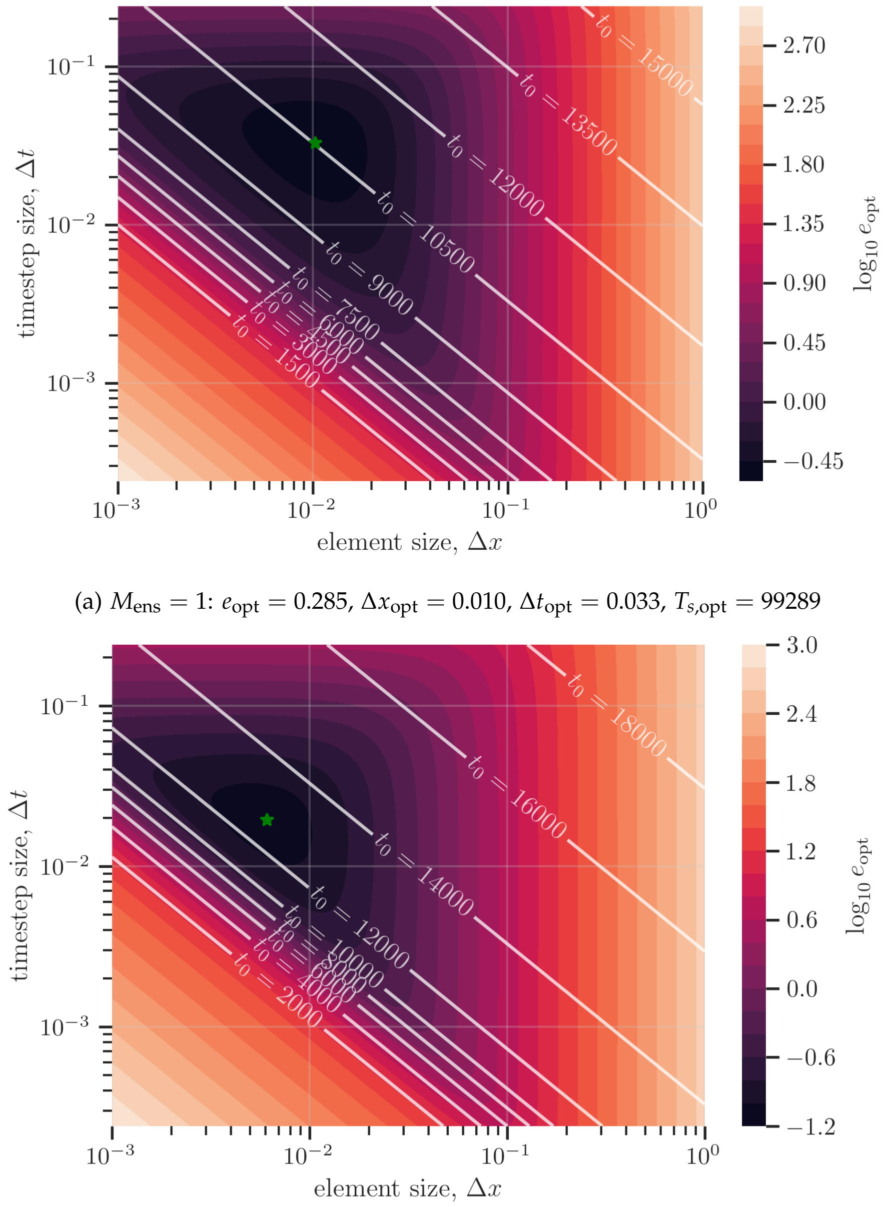
Next, we consider the behavior of the optimizer as a function of computational cost. In Figure 3.15, we can see the convergence behavior for the two choices of discretization used here. From these plots, computed at equivalent budgets according to (3.45), we see two dominant convergence regions emerge, in which

$$e_{\text{opt}} \sim \mathcal{C}_t^{\alpha}. \tag{3.51}$$

When the budget $\mathcal{C}_t$ is large, the convergence behavior is dominated by the sampling behavior and

$$\alpha = -\frac{q_x q_t r}{q_x q_t + q_x r + q_t r}.$$

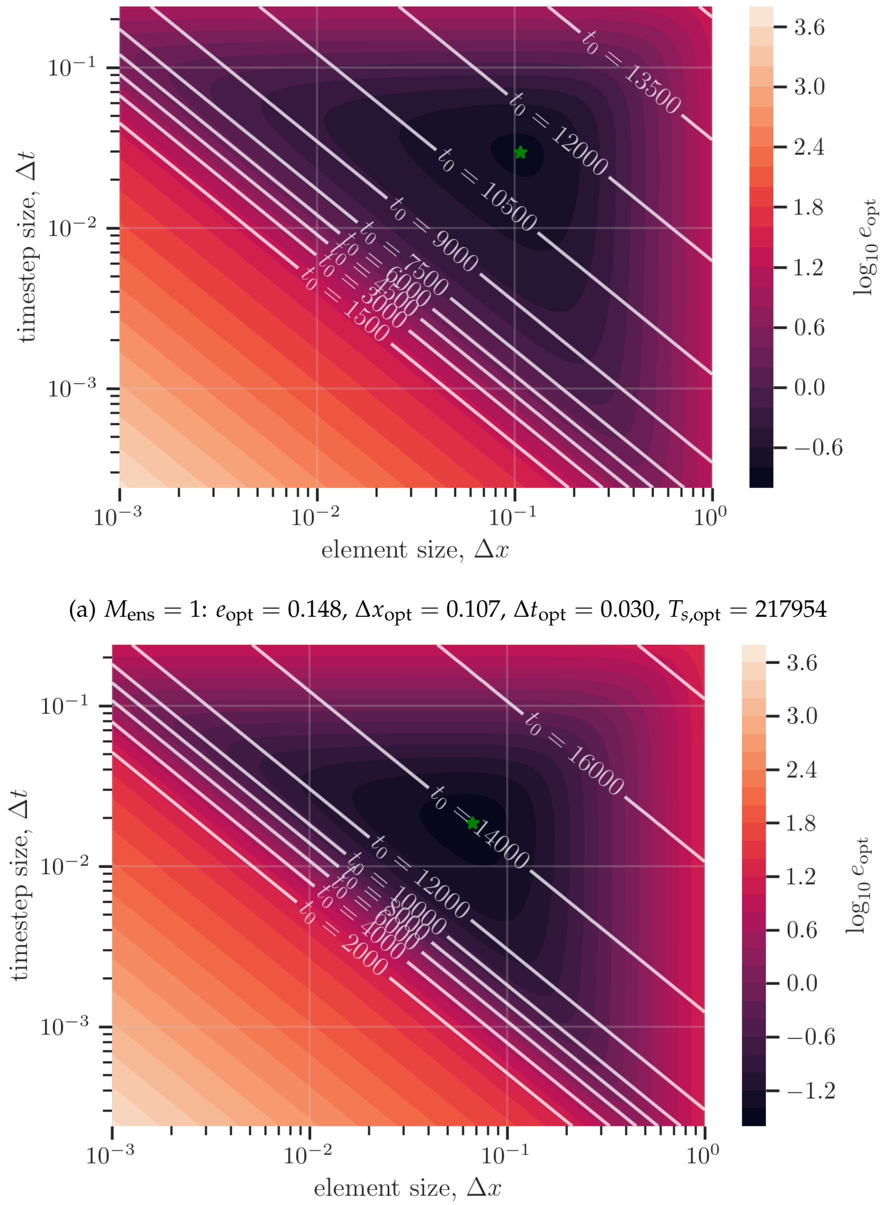
dominates. On the other hand, in the limit of small budgets, the transient resolution becomes a restricting factor. In this region, spin-up costs are the dominant factor, and they are optimally controlled



(a) $M_{\mathrm{ens}} = 1$: $e_{\mathrm{opt}} = 0.285$, $\Delta x_{\mathrm{opt}} = 0.010$, $\Delta t_{\mathrm{opt}} = 0.033$, $T_{s,\mathrm{opt}} = 99289$

(b) $M_{\mathrm{ens}} = 64$: $e_{\mathrm{opt}} = 0.082$, $\Delta x_{\mathrm{opt}} = 0.006$, $\Delta t_{\mathrm{opt}} = 0.019$, $T_{s,\mathrm{opt}} = 25684$

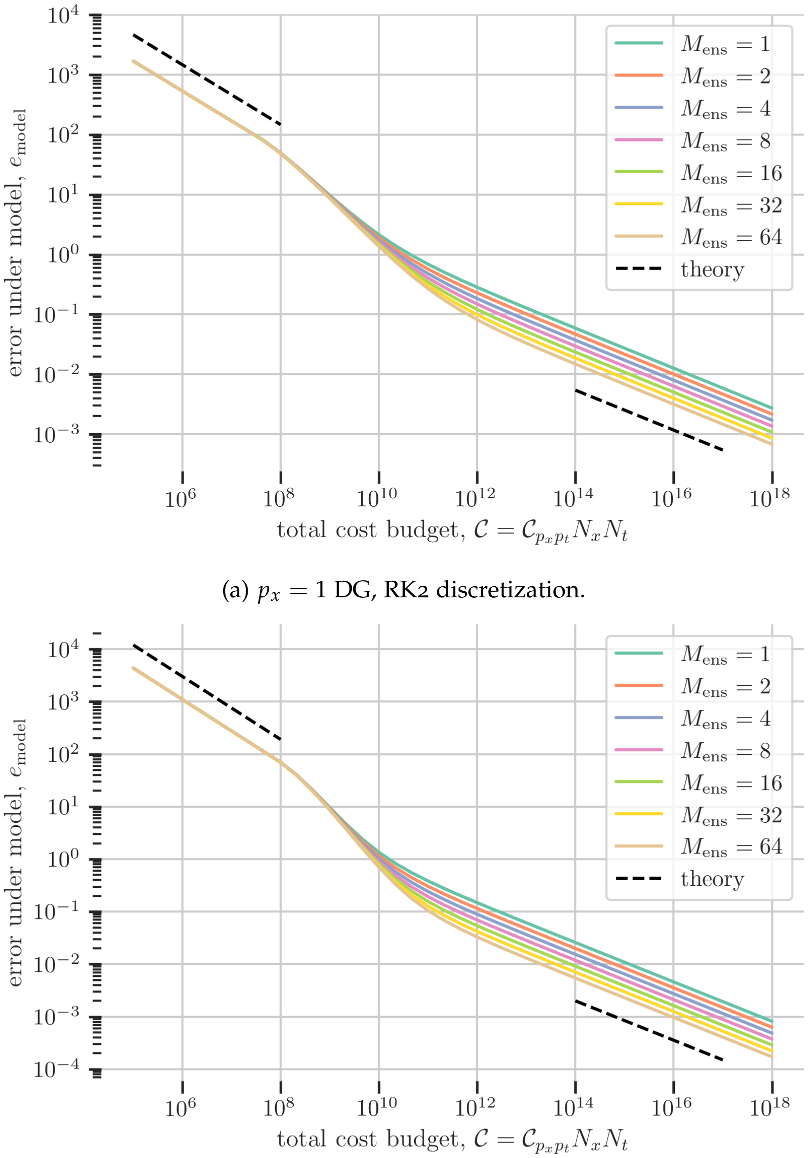
Figure 3.13: Total error model variation with Δx and Δt at optimal choice of t_0 with total element computation budget $N_x N_t = 10^{12}$ using the $p_x = 1/\mathrm{RK2}$ discretization. Contours of optimal choice of t_0 overlaid.



(a) $M_{\text{ens}} = 1$: $e_{\text{opt}} = 0.148$, $\Delta x_{\text{opt}} = 0.107$, $\Delta t_{\text{opt}} = 0.030$, $T_{s,\text{opt}} = 217954$

(b) $M_{\text{ens}} = 64$: $e_{\text{opt}} = 0.033$, $\Delta x_{\text{opt}} = 0.067$, $\Delta t_{\text{opt}} = 0.018$, $T_{s,\text{opt}} = 75353$

Figure 3.14: Total error model variation with Δx and Δt at optimal choice of t_0 with total element computation budget $N_x N_t = 10^{12}$ using the $p_x = 2$/RK3 discretization. Contours of optimal choice of t_0 overlaid.



(a) $p_x = 1$ DG, RK2 discretization.

(b) $p_x = 2$ DG, RK3 discretization.

Figure 3.15: Convergence plots for the expected error under the model at the optimum choice of Δx, Δt, t_0 and T_s at a given wall-clock budget and number of ensembles.

by T_s. Thus, in this region,

$$\alpha = -\frac{q_x q_t}{q_x q_t + q_x + q_t}$$

dominates as long as the discretization errors remain asymptotic. This is the value shown in the left-hand-side theory totems in Figure 3.15.

q_x	q_t	α_{sampling}	α_{spinup}
2	2	$-1/3$	$-1/2$
3	3	$-3/8$	$-3/5$
4	4	$-2/5$	$-2/3$
5	5	$-5/12$	$-5/7$
6	6	$-8/21$	$-3/4$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
∞	∞	$-1/2$	-1

Table 3.6: Total error convergence rates as a function of design order under the model.

In Table 3.6, we show these theoretical cost-error rates for discretizations with $q_x = q_t$. As seen for ODEs in Section 1.2, the optimal rates achievable with a higher-order method on a chaotic system are limited by the central limit theorem. Meanwhile, for problems with significant spin-up transient costs, the rate for small budgets is limited by T_s^{-1}. Finally, we note that it appears the high-order rates can be recovered in the intermediate region between statistically dominated rates, between 10^7 and 10^9 cost units in Figure 3.15. It is possible that for some problems these properties might open up a significant region with high-order convergence with respect to $\mathcal{C}$, but it is certainly has an insignificant impact in this setting.

3.6 Small-sample model identification

In order to demonstrate and validate the Bayesian procedure, we will now perform small-sample fits. To do so, a database of $J_{T,hp}$ is generated. We generate random choices of Δx and Δt subject to a sampling budget of $\mathcal{C}_s = 10^7$, by choosing Δx and Δt sampled from a loguniform distribution such that:

$$\log_{10} \Delta x \sim \mathcal{U}(\log_{10} \Delta x_{\text{max}} - 1, \log_{10} \Delta x_{\text{max}})$$
$$\log_{10} \Delta t \sim \mathcal{U}(\log_{10} \Delta t_{\text{max}} - 1, \log_{10} \Delta t_{\text{max}})$$

then T_s is set by the budget, as in (3.15). The maximum values of Δx and Δt are set using the limits in Table 3.1. Sampled triples for which T_s is greater than the minimum for convergence are rejected; moreover, the system is spun up for $t_0 = 12{,}000$ and samples for which the total cost $\mathcal{C} > 10^{10}$ are also rejected. The errors with respect to J_{ref} in the resulting set of simulations for $p_x = 1$, RK2 simulation– on which we will concentrate in this section– can be found in Figure 3.16.

For each forthcoming Bayesian fit, we will bootstrap a length-M sets of simulation results, $\{J_{T,hp}\}$, out of the database. Then, $\{J_{T,hp}\}$, (3.35), (3.36), and (3.41) are used to compute the MAP estimate θ^*_{MAP} as well as an estimate of θ_{MAP} and the error models using (3.33). In Figure 3.17, the first such Bayesian fit is shown, with $M = 1000$

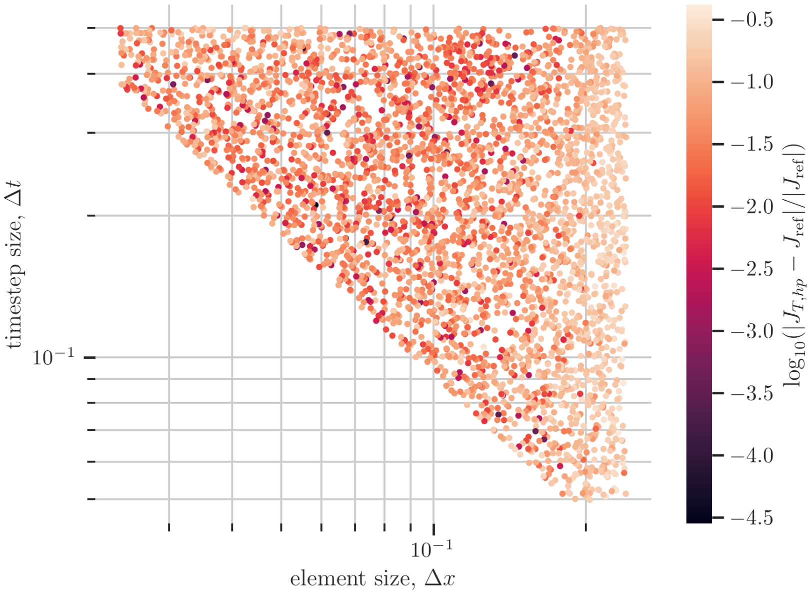


Figure 3.16: Error values of output data for $\mathcal{C}_s = 10^7$ and $t_0 = 12{,}000$. Area of markers scale with T_s subject to budget.

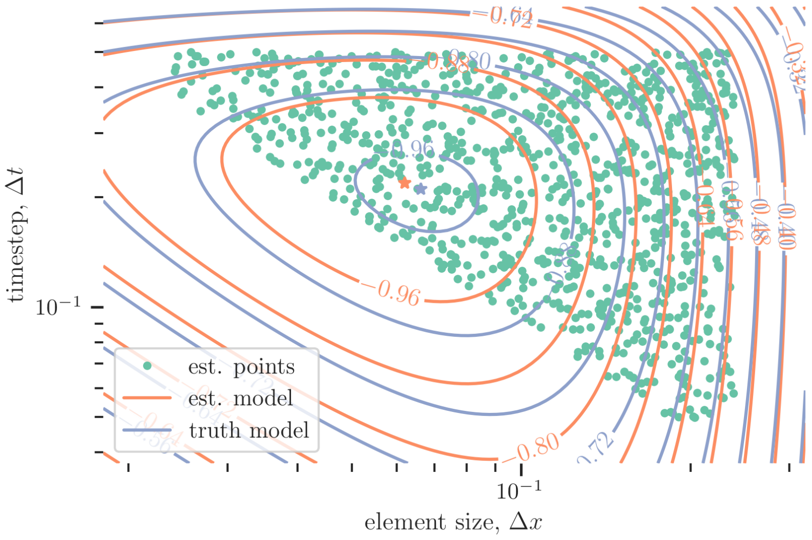


Figure 3.17: Bayesian error model fit with $M = 1000$ simulation datapoints drawn randomly. Small-sample estimate shown with white contours, "true" model shown with black and colored background contours.

points. This result is a fit that captures the important qualitative features, namely the location and magnitude of error the minimizer Quantitatively, the fit has a small mismatch with the true value of the optimal error, and a slight offset of the optimizer. The values and their comparison to the values in Table 3.2 are given in Table 3.7.

variable	value	pct. error
$(C^*_{p_x})_{\mathrm{MAP}}$	478.23	0.46%
$(B^*_{p_t})_{\mathrm{MAP}}$	-40.69	14.41%
$(A^*_0)_{\mathrm{MAP}}$	59.57	16.77%
$(J_\infty)_{\mathrm{MAP}}$	118.75	0.26%
variable	value	ref. value
$(\Delta x_{\mathrm{opt}})_{\mathrm{MAP}}$	0.062	0.066
$(\Delta t_{\mathrm{opt}})_{\mathrm{MAP}}$	0.211	0.211
$(e_{\mathrm{opt}})_{\mathrm{MAP}}$	10.911	12.665

Table 3.7: Small-sample fit results for $p_x = 1$, RK2 with $M = 1000$.

These results show that the error model captures J_∞ and $C^*_{p_x}$ under 5% error, while the temporal and sampling error models have larger errors. To further interrogate these results, we can look at the convergence behavior as understood by the model in Figure 3.18. These results show that temporal-discretization-dominated simulations are relatively few, compared to the spatial-discretization- and sampling-dominated simulations. In addition to the lack of sample density across the error effects, it is likely that non-asymptotic sampling behavior can be present when $T_s \lesssim 30$, based on the results in Section 3.4.

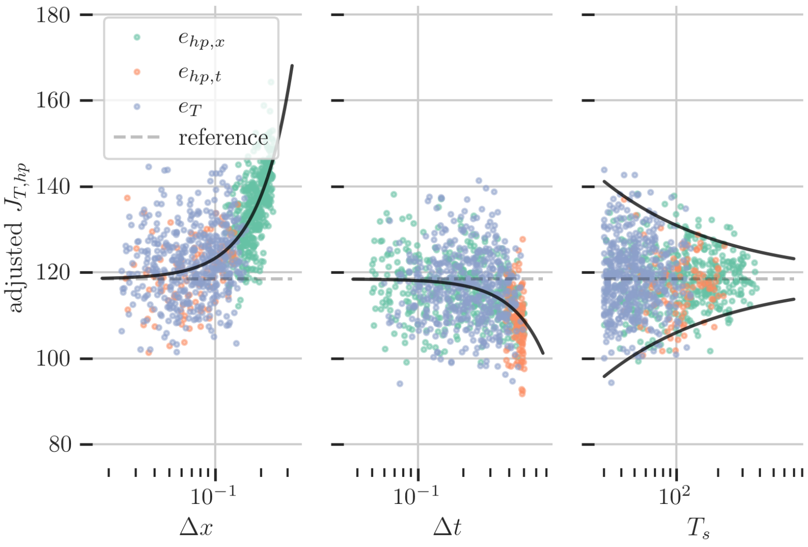


Figure 3.18: Convergence behavior under small-sample fit for $p_x = 1$, RK2 with $M = 1000$. Adjusted $J_{T,hp}$ removes modeled temporal effects for spatial convergence, vice versa, and removes both spatial and temporal effects for the sampling error convergence. Data colored by dominant effect under observed model.

In addition to the results shown here, we have also tested the Bayesian small sample procedure with synthetic data generated exactly according to the likelihood function, for which this quantiative mismatch does not exist as $M \to \infty$. This suggests that the real simulation data containts some non-asymptotic effects, whether in Δx, Δt, or T_s or due to uncontrolled initial transients. The synthetic data studies can be found in Appendix D.

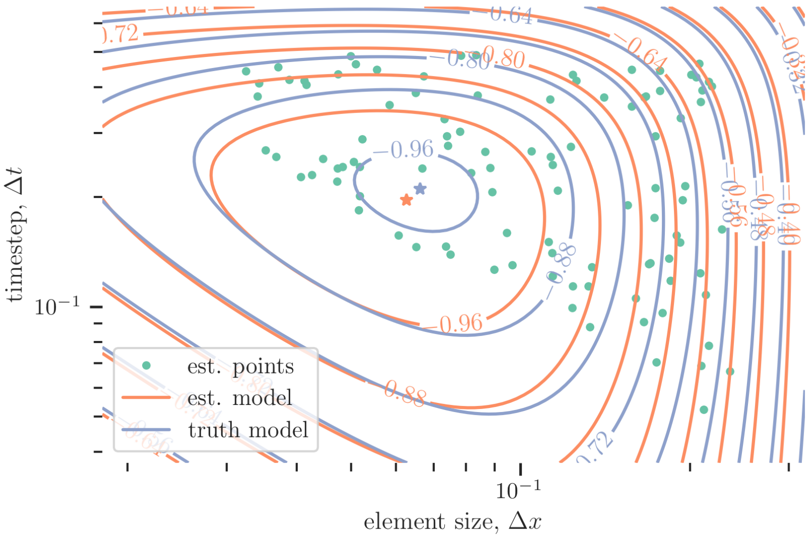


Figure 3.19: Bayesian error model fit with $M = 100$ simulation datapoints drawn randomly. Small-sample estimate shown with white contours, "true" model shown with black and colored background contours.

variable	value	pct. error
$(C^*_{p_x})_{\text{MAP}}$	479.51	0.19%
$(B^*_{p_t})_{\text{MAP}}$	−43.33	8.86%
$(A^*_0)_{\text{MAP}}$	55.65	22.25%
$(J_\infty)_{\text{MAP}}$	119.88	1.22%
variable	value	ref. value
$(\Delta x_{\text{opt}})_{\text{MAP}}$	0.061	0.066
$(\Delta t_{\text{opt}})_{\text{MAP}}$	0.201	0.211
$(e_{\text{opt}})_{\text{MAP}}$	10.542	12.665

Table 3.8: Small-sample fit results for $p_x = 1$, RK2 with $M = 100$.

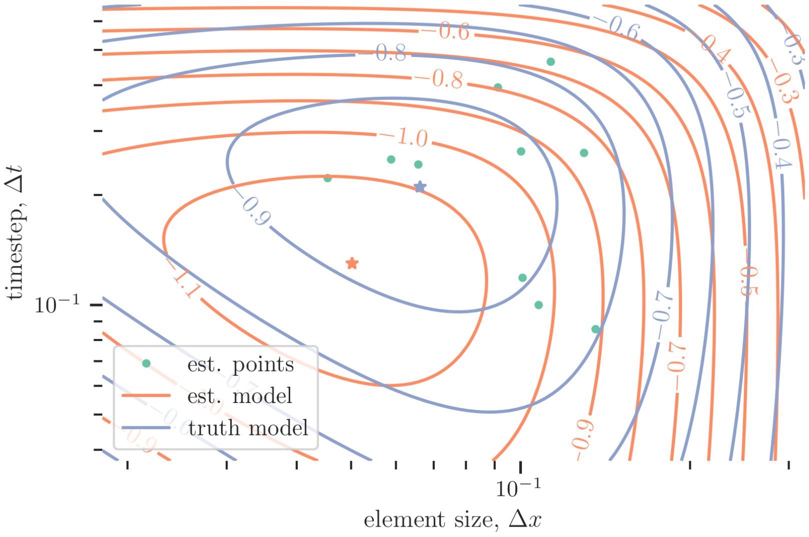


Figure 3.20: Bayesian error model fit with $M = 10$ simulation datapoints drawn randomly. Small-sample estimate shown with white contours, "true" model shown with black and colored background contours.

variable	value	pct. error
$(C^*_{p_x})_{\text{MAP}}$	465.70	3.06%
$(B^*_{p_t})_{\text{MAP}}$	−41.72	12.24%
$(A^*_0)_{\text{MAP}}$	36.20	49.42%
$(J_\infty)_{\text{MAP}}$	118.59	0.12%
variable	value	ref. value
$(\Delta x_{\text{opt}})_{\text{MAP}}$	0.053	0.066
$(\Delta t_{\text{opt}})_{\text{MAP}}$	0.177	0.211
$(e_{\text{opt}})_{\text{MAP}}$	7.826	12.665

Table 3.9: Small-sample fit results for $p_x = 1$, RK2 with $M = 10$.

In Figures 3.19 and 3.20, the fit results for $M = 100$ and $M = 10$ simulations are shown, respectively, with the corresponding data in Tables 3.8 and 3.9. These fits again identify J_∞ with less than 1% error while having larger error for other model parameters. Nonetheless, these anecdotal results for $M = 1000$, 100, and 10 demonstrate that the small-sample procedure can identify J_∞ and make high-quality qualitative estimates of the optimal cost-constrained discretization $(\Delta x_{\text{opt}}, \Delta t_{\text{opt}})$, even approaching the small-sample limit.

Finally, we study the sensitivity of the small-sample procedure to the particular set of M simulations of $J_{T,hp}$. In order to do so, we will now repeat the process of the small-sample fits over $\mathcal{M} = 100$ length-M sets. Each of these sets will be bootstrapped without replacement from the database of simulations used previously in this section, without replacement at a given $\mathcal{M}$. Due to computational limits, $\mathcal{M} = 100$ is not available for every choice of M. Figure 3.21 shows the $\mathcal{M}$ available in the data we have. The following plots will have $\mathcal{M}$ according to this limited data.

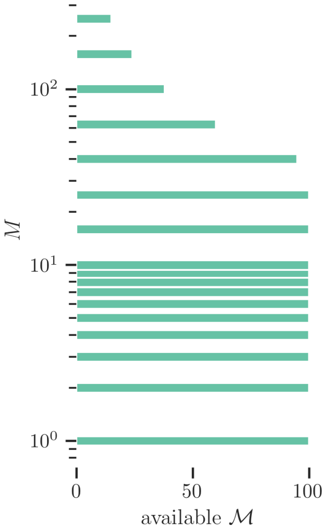
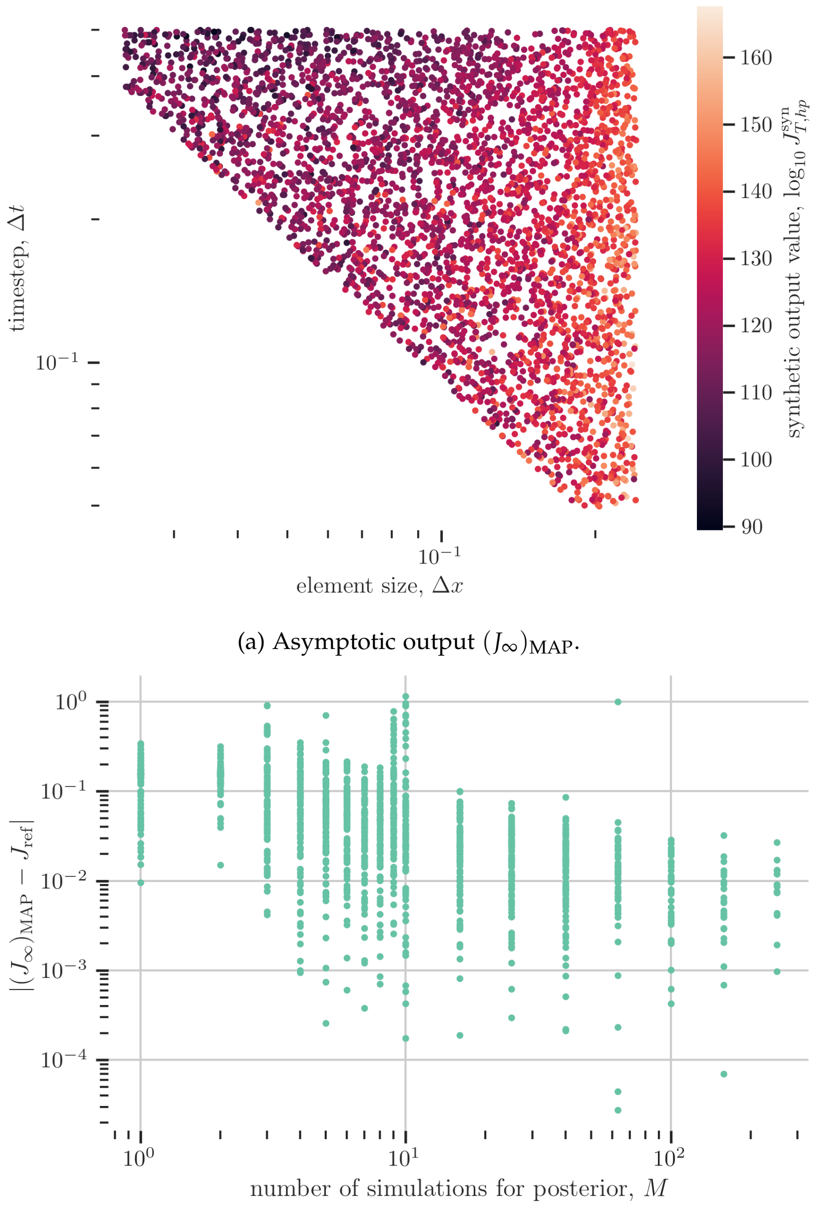


Figure 3.21: $\mathcal{M}$ available with computed data.

In Figure 3.22, the $(J_\infty)_{\mathrm{MAP}}$ estimates are shown. This plot demonstrates that as M increases, the distribution of $(J_\infty)_{\mathrm{MAP}}$ converges towards J_{ref} with a rate approximately $M^{-1/2}$. In Figure 3.23, we consider the accuracy of error estimates and optimized discretizations that are generated by each of the length-M MAP estimates. In these plots, we see the quality of the error model, measured by the approach of $(e_{\mathrm{opt}})_{\mathrm{MAP}}$ to $(e_{\mathrm{opt}})_{\mathrm{ref}}$, is good and stays within a small factor of the reference error. In the lower subplot of Figure 3.23, we show the multiplicative factor between the error under the reference model at $(e_{\mathrm{opt}})_{\mathrm{MAP}}$ and the optimal error using the reference model, which should represent the lowest possible error at this cost. This plot shows that, using the small sample estimation capability here, more than 90% of the estimates with random length-M sets $\{J_{T,hp}\}$ with $M \geq 10$ should result in a simulation with no more than a factor of two more than the minimum possible error.

3.7 *Conclusions and future work*

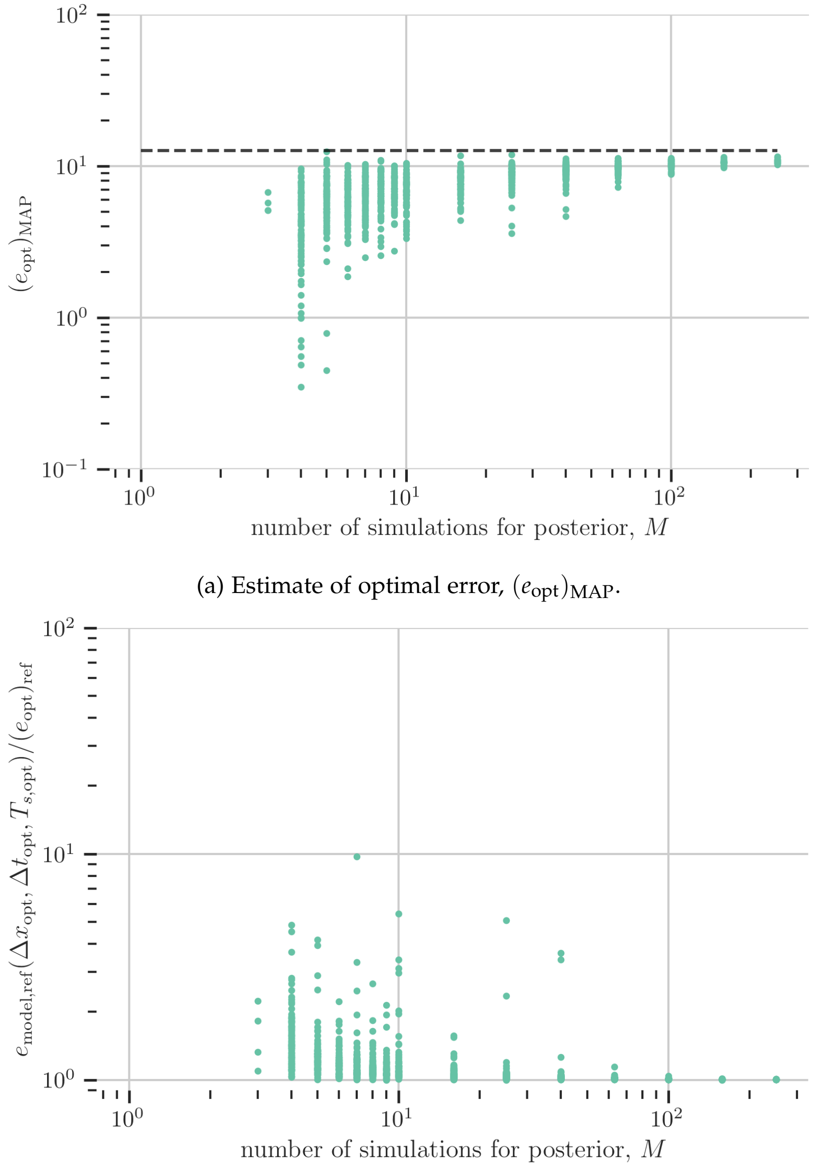
The results in this chapter demonstrate that the error model can be extended to the domain of chaotic PDEs and has explanatory value in that context. Moreover, we demonstrate that these effects can be identified with a small-sample approach without relying on the use of expensive reference computations. In addition to the sampling behavior, we show that the spin-up transient behavior of the PDE systems, as in the ODE case, can be identified and accounted for. Taken together, we can now identify all of the contributions to the error in $J_{T,hp}$, without reference computations that far exceed the cost of $J_{T,hp}$.



(a) Asymptotic output $(J_\infty)_{\mathrm{MAP}}$.

(b) Error $|(J_\infty)_{\mathrm{MAP}} - J_{\mathrm{ref}}|$.

Figure 3.22: Small sample identification of J_∞ as a function of M.



(a) Estimate of optimal error, $(e_{\mathrm{opt}})_{\mathrm{MAP}}$.

(b) Excess error in optimized $(M+1)$-th simulation.

Figure 3.23: Small sample error estimation as a function of M. In (a): 295 total points out of range: 100 at $M = 1$, 100 at $M = 2$, 95 at $M = 3$. In (b): 300 total points out of range: 100 at $M = 1$, 100 at $M = 2$, 95 at $M = 3$, 1 at $M = 6$, 1 at $M = 9$, 2 at $M = 10$, 1 at $M = 40$.

4

Bayesian optimization of discretizations for ergodic chaotic differential equations

Start where you are. Use what you have. Do what you can.

—Arthur Ashe

As we have shown in Chapter 3, it is possible to estimate the transient behavior of a discretized PDE. Additionally, we can identify a model for the discretization and sampling errors of a given physical system and discretization using a small number of outputs from simulations measured on the attractor. Now, we turn towards application of these results *in situ* when running simulations without a priori knowledge of the error model or transient behavior.

4.1 Small-sample identification with non-negligible transient effects

In Section 3.6, we developed a small-sample approach for the PDE problem when $t_0 \gg T_\lambda$. The first key problem in this chapter is that we will no longer be able to assume that $t_0 \gg T_\lambda$ and that the spin-up transient is negligible, since the target error tolerance is not known a priori. In (3.24), we developed a model for the expected value of a simulation with discretization and statistical errors– including both sampling and spin-up errors– by assuming an exponential decay model for the convergence to the solution on the attractor, (1.52), and using its expected integrated value. In the small-sample case, we will assume that the coefficients in (1.52) can be estimated for a given simulation. Now, we can insert (1.52) into (3.8), and follow the derivation in Section 3.6 to arrive at a likelihood function for a simulation result $J_{T,hp}$ subject to a spin-up transient:

$$p\left(\left\{J_{T,hp}^{(i)}\right\} \middle| J_\infty, C_{p_x}^*, q_x, B_{p_t}^*, q_t, A_0^*, r\right) = \prod_{i=1}^{N_{\text{samp}}} \mathcal{N}\left(J_{T,hp}^{(i)}; J_\infty + C_{p_x}^* \Delta x^{q_x} + B_{p_t}^* \Delta t^{q_t} + e_\lambda^{(i)}, \left(A_0^* T_s^{-r}\right)^2\right), \quad (4.1)$$

where

$$e_\lambda^{(i)} \equiv A_\lambda^{(i)} \frac{T_\lambda^{(i)}}{T_s^{(i)}} \exp\left(-\frac{t_0^{(i)}}{T_\lambda^{(i)}}\right) \left[1 - \exp\left(-\frac{T_s^{(i)}}{T_\lambda^{(i)}}\right)\right] \quad (4.2)$$

gives the error contribution based on the i-th $g_{hp}^{(i)}$ signal, which is used to generate $J_{T,hp}^{(i)}$.

Computational problem

In this chapter, we will develop an algorithmic approach that is compatible with the PDE models derived in Chapter 3. However, in order to moderate computational costs, we will simulate the Kuramoto-Sivashinsky equation as in Section 3.3 using a fixed number of spatial degrees of freedom. In this case, we fix the number spatial elements at $N_x = 64$ and $p_x = 2$, in which event we will refer to the system as the KSE-ODE system. This alludes to the fact that the result reduces to an ODE problem, as in Chapter 1 and 2, but we will refer to the PDE-specific equations in Chapter 3 to retain generality, assuming that $C_{p_x}^* \to 0$ everywhere in the results for the KSE-ODE case.

4.2 *Numerical results: error model identification in the presence of spin-up transient effects*

In this section, we consider the error in a KSE-ODE simulation. Because we have fixed the spatial system far from the convergence region, we will not recycle any of the reference results used previously, as there are no guarantees that the temporal behavior of the spatially coarse system will be equivalent. We compute a reference simulation with an ensemble of $M_{\text{ens}} = 25$ instances, with $T_s = 2.73 \times 10^5$, $t_0 = 3577$, and $\Delta t = 2.67 \times 10^{-2}$. The result is

$$J_{\text{ref}} \approx 181.30 \pm 0.04, \quad (4.3)$$

where the limits given are based on the standard error on the ensemble of averages.

To elaborate on this point: multiscale temporal behaviors are dependent on the physical mechanisms resolved, so when the physical system is discretized with few spatial elements we can expect the temporal behavior to be significantly distinct. Grid-dependence of the temporal properties of CFD problems, for instance, is very clearly demonstrated in Fernandez & Wang [2017].

Now we can take the likelihood function of a simulation result $J_{T,hp}$ and subtract J_∞ to find:

$$p\left(e_{T,hp}\right) = \mathcal{N}\left(J_{T,hp} - J_\infty;\, B^*_{p_t}\Delta t^{q_t} + e_\lambda,\, \left(A^*_0 T_s^{-r}\right)^2\right). \tag{4.4}$$

which can be manipulated analogously to (3.31) to allow the transformation in (3.33) and equivalence to (3.24). Now, in order to complete the process in Section 3.4 with the transient correction, we must treat the e_λ term.

In this section, we fit a function μ_g

$$\mu_g^{(i)} = J_{\infty,hp}^{\text{decay},(i)} + A_\lambda^{(i)} \exp\left(-\frac{t}{T_\lambda^{(i)}}\right) \tag{4.5}$$

to $g_{hp}^{(i)}(t)$ using a non-linear least squares procedure, in order to find $A_\lambda^{(i)}$ and $T_\lambda^{(i)}$, with $J_{\infty,hp}^{\text{decay},(i)}$ used to denote the asymptotic value as detected by the mean fit process. With this estimate, we can compute $e_\lambda^{(i)}$ using (4.2). Applying this correction, we can then use the large-sample method of Section 3.4 to find the error model parameters in (4.4).

The (asymptotically equivalent) Bayesian least squares procedure used to fit the decay in Chapter 1 can be prone to ill-fitting for the KSE-ODE problem, hence the use of standard nonlinear least-squares.

To assess the correction method and identify the various error model parameters, we will simulate using $N_s = 4.0 \times 10^4$, various selections of t_0, and the RK3 discretization, with Δt sampled loguniformly subject to the convergence limits on Δt and T_s in Table 3.1. In addition to the minimum on T_s, we also require that $T = t_0 + T_s \geq 1200$, in order to promote accurate capture of the transient behavior, and simulations for which the nonlinear least squares fit fails on to capture ψ^* are rejected. To establish a reference estimate of the parameters, we start with 2000 simulations with $t_0 = 9000$. At this large value of t_0, e_λ is effectively negligible, with the 99.7th-percentile value of $|e_\lambda|$ approximately 5.9×10^{-2}. In Table 4.1, the values of the resulting fit are shown, which we take as the canonical reference estimates of the system's error behavior (i.e. θ^* and ψ).

$B^*_{p_t}$	q_t	A^*_0	r	J_∞	$(\lvert A_\lambda \rvert)_{95}$	$(T_\lambda)_{95}$
-311.3	3	115.7	1/2	181.34	126.12	1111.34

Table 4.1: Results of error model fit for $p_x = 2$, RK3 at $N_s = 4 \times 10^4$ with error correction at $t_0 = 9000$. Simulations sampled log-uniform in Δt with $\Delta t < 0.24$ and $T_s > 9580.0$; 1835 simulations used for fit.

Now, we can repeat the study with lower values of t_0; for an extremal case, we now explore simulations with $t_0 = 300$, but otherwise specifying simulations identically to the previous case. For this case, we simulate 1200 runs, truncating the range such that $T_s > 2400$. In Figure 4.2, we show the raw data that results. Figure 4.2

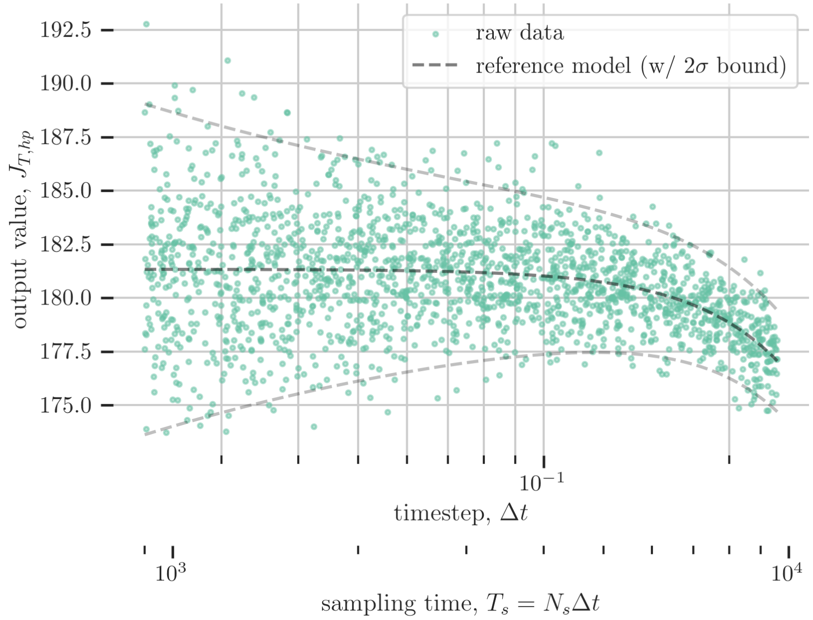


Figure 4.1: Raw $J_{T,hp}$ data from $N_x = 64$, $p_x = 2$ simulation using RK3 with $t_0 = 9000$ (reference case). Fit superimposed.

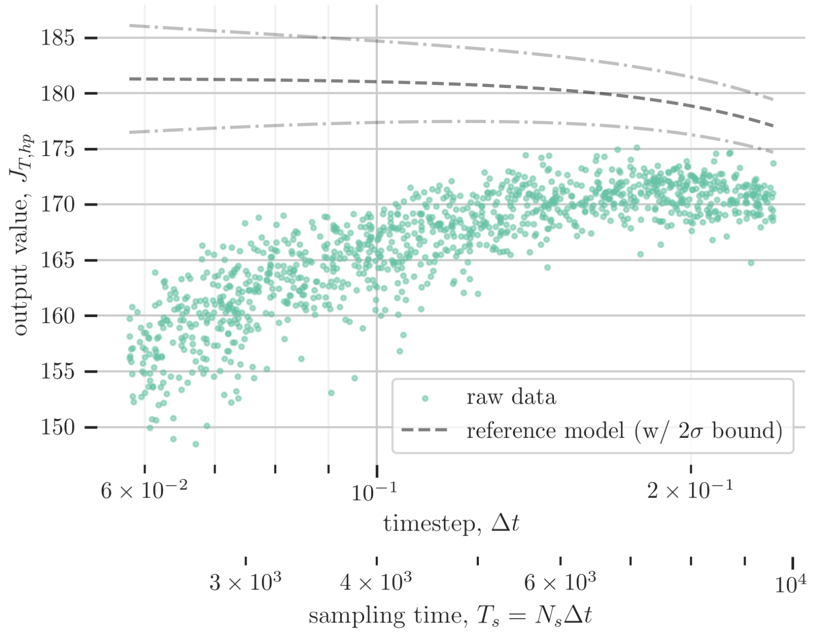


Figure 4.2: Raw $J_{T,hp}$ data from $N_x = 64$, $p_x = 2$ simulation using RK3 with $t_0 = 300$. Reference model fits superimposed.

demonstrates behavior that clearly departs from the transient-free results from the previous chapters, such as Figure 2.11. Specifically, there exists a clear trend as Δt shrinks, away from J_∞-centered sampling.

This behavior emerges due ot the effect of the transient error. In Figure 4.3, we have estimated the transient error $e_\lambda^{(i)}$, each calculated with the nonlinear-least squares estimate of the decay computed with the weighted least-squares fit to $g_{hp}^{(i)}$. We expect the T_s-dependent behavior of the transient effect to scale with T_s^{-1}, as in (1.52). This behavior is reflected in both the mean and statistical variation of e_λ, which can be clearly seen in Figure 4.4. On one hand, the trend in Figure 4.4 suggests that the estimates of e_λ match the theoretically expected behavior with T_s. On the other hand, these plots– especially Figure 4.3– show how challenging it is to model the transient effect, which can materialize as a random effect whose non-zero mean value grows with T_s^{-1} as T_s shrinks, while also having the same trend in its standard deviation.

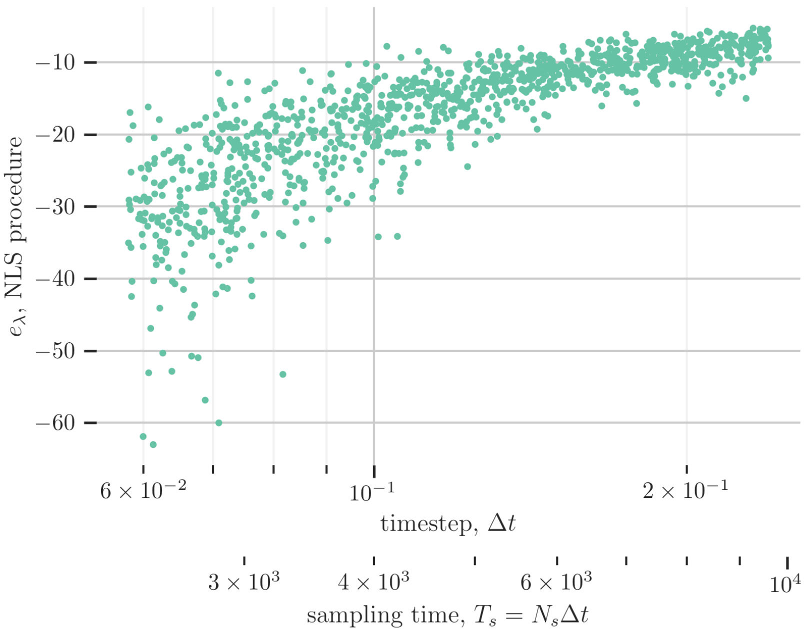


Figure 4.3: Estimates of e_λ for $J_{T,hp}$ data from $N_x = 64$, $p_x = 2$ simulation using RK3 with $t_0 = 300$.

Given that the values of e_λ converge as expected, we now apply e_λ as a correction to the $J_{T,hp}$ data, in order to attempt to recover the underlying discretization and sampling error model. We first compute $e_\lambda^{(i)}$ using the method above. Then, we can repeat the nonlinear least

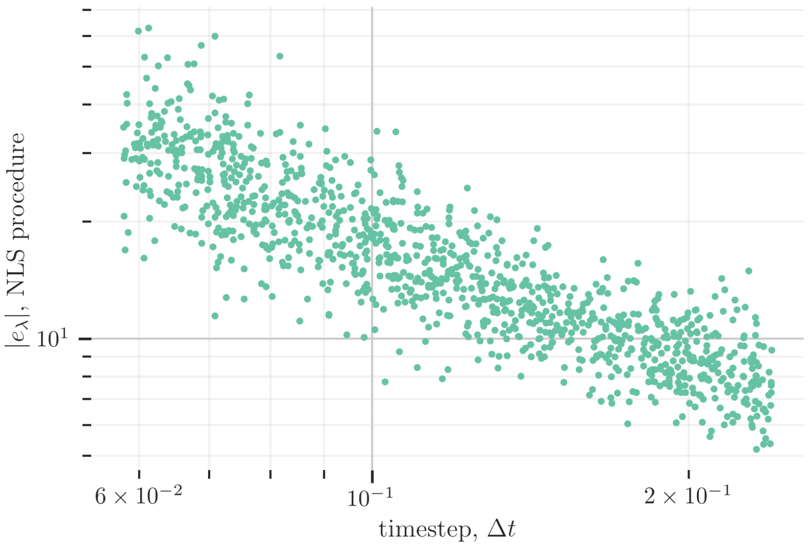


Figure 4.4: Estimates of $|e_\lambda|$ for $J_{T,hp}$ data from $N_x = 64$, $p_x = 2$ simulation using RK3 with $t_0 = 300$.

squares procedure, using an augmented mean function

$$J_\infty + C^*_{p_x}(\Delta x^{(i)})^{q_x} + B^*_{p_t}(\Delta t^{(i)})^{q_t} + e^{(i)}_\lambda,$$

about which $J^{(i)}_{T,hp}$ should me sampled with a standard deviation that scales with $T_s^{-1/2}$. We fit for $B^*_{p_t}$, A^*_0, and J_∞ using (4.1) as in Section 3.4.

In Figure 4.5 we report the error under the resulting error model. Here, the qualitative behavior that we expect on the attractor is present, but there are clear effects that are not described by the central limit theorem behavior for $T_s < 4000$, presumptively induced by the e_λ corrections. Despite the inaccuracies, Figure 4.5 allows significantly more insight into the behavior of the system than the equivalent raw data in Figure 4.2. As in Section 3.4– but here adding the effect of e_λ– we expect the quantity

$$\frac{J_{T,hp} - e_\lambda - J_\infty}{T_s^{-1/2}}$$

to follow a normal distribution; in Figure 4.6, we show that the result of the nominally normally-distributed quantity after transient correction does appear to be approaching a normal distribution, although not without some evident defects due to error in the transient correction. In Table 4.2, the resulting reference values from the fit for $B^*_{p_t}$ and A^*_0 and J_∞ with q_t and r fixed at their asymptotic values are given, along with conservative (95% percentile) estimates of T_λ

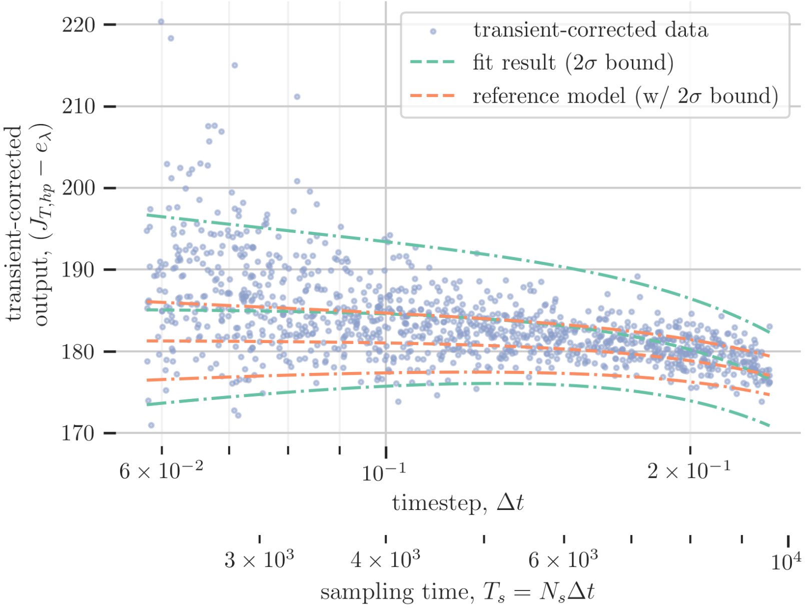


Figure 4.5: Corrected $J_{T,hp}$ data from $N_x = 64$, $p_x = 2$ simulation using RK3 with $t_0 = 300$.

$B^*_{p_t}$	q_t	A^*_0	r	J_∞	$(\lvert A_\lambda \rvert)_{95}$	$(T_\lambda)_{95}$
-624.7	3	279.5	1/2	185.20	127.16	1194.00

Table 4.2: Results of error model fit for $p_x = 2$, RK3 at $N_s = 4 \times 10^4$ with error correction at $t_0 = 300$. Simulations sampled log-uniform in Δt with $\Delta t < 0.24$ and $T_s > 9584.0$; 1115 simulations used for fit.

and $|A_\lambda|$. We note from the values in Table 4.2 that the transient correction also appears to induce an offset in J_∞ with respect to J_{ref}.

For comparison, we also have run 1600 simulations with $t_0 = 1000$ and $t_0 = 3000$ but otherwise identical to further investigate the quality of the spin-up transient correction. In Figures 4.7 and 4.8 the results of this study are shown, and corresponding data is given in Tables 4.3 and 4.4.

$B^*_{p_t}$	q_t	A^*_0	r	J_∞	$(\lvert A_\lambda \rvert)_{95}$	$(T_\lambda)_{95}$
-647.8	3	250.8	1/2	185.46	127.73	1221.60

Table 4.3: Results of error model fit for $p_x = 2$, RK3 at $N_s = 4 \times 10^4$ with error correction at $t_0 = 1000$. Simulations sampled log-uniform in Δt with $\Delta t < 0.24$ and $T_s > 9598.0$; 1461 simulations used for fit.

These results show that t_0-dependence in the quality of small-sample fits becomes a factor in the fits when $J_{T,hp}$ is corrected for the e_λ contribution. Particularly, small t_0 leads to excess variance induced by the correction term, and causes drift in the estimation of J_∞. Nonetheless, fits to the corrected data give a plausible qualitative approximation of the real error behavior even with significant e_λ when the transient is captured by sufficient $T = t_0 + T_s$, in spite of evident spin-up effects. This qualitative success should be sufficient to allow us to proceed with the assimilation of e_λ-corrected data and

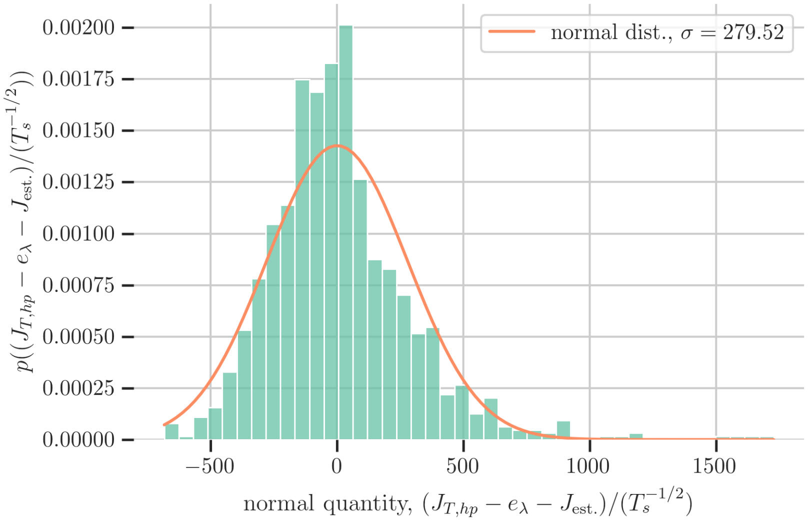


Figure 4.6: Histogram of nominally normal quantity from $J_{T,hp}$ data from $N_x = 64$, $p_x = 2$ simulation using RK3 with $t_0 = 300$.

$B^*_{p_t}$	q_t	A^*_0	r	J_∞	$(\lvert A_\lambda \rvert)_{95}$	$(T_\lambda)_{95}$
−371.3	3	134.5	1/2	182.39	125.34	1125.33

Table 4.4: Results of error model fit for $p_x = 2$, RK3 at $N_s = 4 \times 10^4$ with error correction at $t_0 = 3000$. Simulations sampled log-uniform in Δt with $\Delta t < 0.24$ and $T_s > 9598.0$; 1494 simulations used for fit.

make estimates of the discretization and sampling errors using the small-sample process.

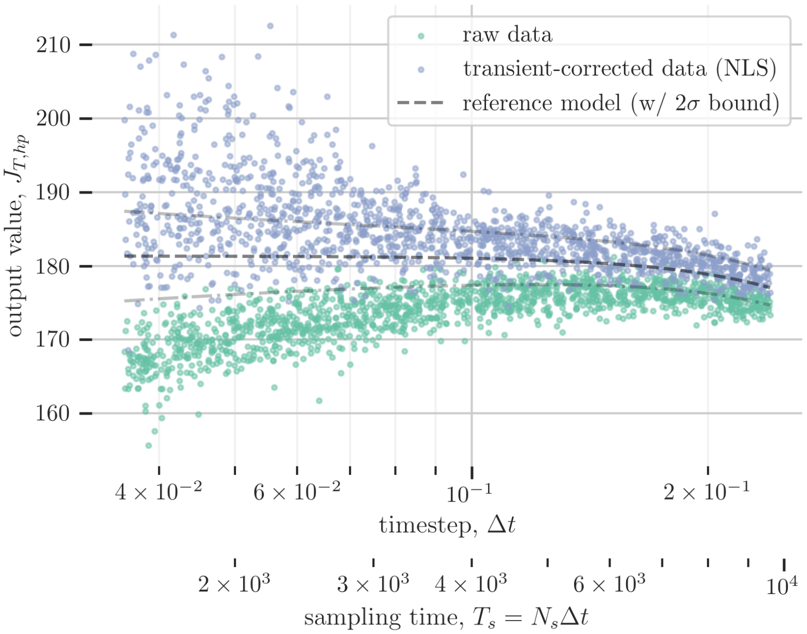


Figure 4.7: Raw and corrected $J_{T,hp}$ data from $N_x = 64$, $p_x = 2$ simulation using RK3 with $t_0 = 1000$.

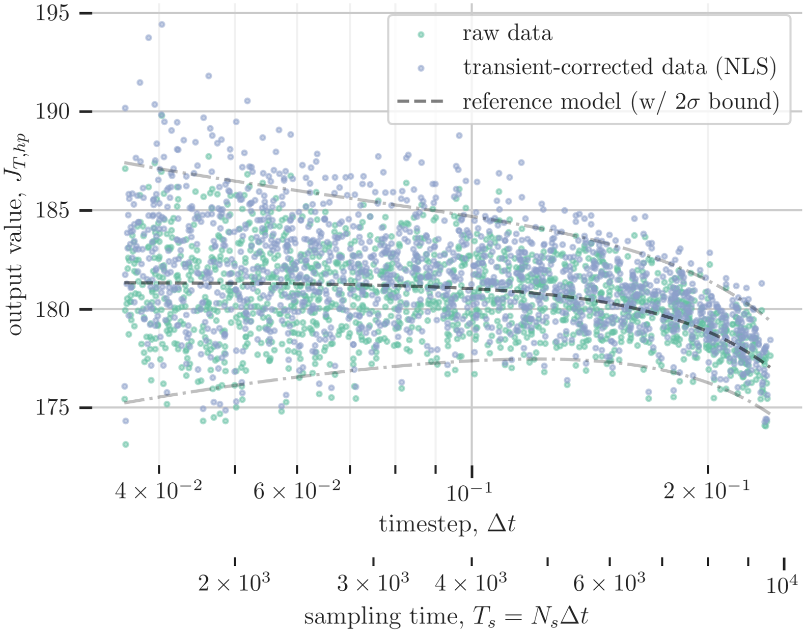


Figure 4.8: Raw and corrected $J_{T,hp}$ data from $N_x = 64$, $p_x = 2$ simulation using RK3 with $t_0 = 3000$.

4.3 System analysis of simulation components

Having demonstrated that the use of a transient correction allows us to recover useful approximations of the attractor model, we turn toward demonstrating that the result of Chapters 2 and 3– model identification at small sample sizes– can be recreated for simulations with spin-up transients. Now, we want to combine these results and create a scheme to simulate optimally with respect to *total* cost without a priori knowledge of the error or solution behavior. In order to understand the requirements of such a scheme, we now summarize and consider the interconnections between the inputs and outputs of the various estimation processes we have thus far developed.

Any given simulation, which we will denote symbolically by $\mathrm{PDE}_{hp}(\cdot)$, takes as inputs gridsize Δx, timestep Δt, and total simulation time T. The result of this simulation is the state of the system as a function of space (if applicable) and time, $\mathbf{u}_{hp}(\vec{x}, t)$. In addition to the state, an instantaneous output quantity $g_{hp}(t)$ is computed as well.

Average values of outputs, $J_{T,hp}$ are computed by integrating the output signal $g_{hp}(t)$ starting at some specified time t_0 through the remaining T_s. In addition to output integration, g_{hp} is also used for the decay fitting process, denoted symbolically by $\mathrm{decay}(\cdot)$. This process results in estimates of the observed transient behavior, ψ^*. The relation (3.21) between the choice of t_0 and T_s and the amount of spin-up error can be derived from ψ^*.

Last but not least, the small-sample fitting process, takes a set of output quantities $\{J_{T,hp}\}$ and gives a set of model parameters that describes the sampling and discretization error contributions, θ^*. These, alongside the decay behvaior ψ^*, determine the optimal choices of Δx, Δt, t_0, and T_s for at a given cost, at which $\mathrm{PDE}_{hp}(\cdot)$ should be run.

Unfortunately, this analysis makes clear the challenge posed by the framework. In an optimal situation, any given simulation should be situated at an optimal choice of $(\Delta x, \Delta t, T_s)$ under (3.27). This requires knowledge of both ψ^* and θ^*. Once a simulation is complete, ψ^* can be estimated directly by the decay fit but before the simulation is run, the quantity is not known and requires estimation. Meanwhile, θ^* is in general not known but we aim to estimate it using a small set of sampled simulations, $\{J_{T,hp}\}$. Calculation of these average outputs $J_{T,hp}$ can be done with the output of any given simulation, but nominally, we want an optimal balance of spin-up and sampling, subject again to (3.27). However, this optimal integration must be performed Δx and Δt fixed, since the simulation has already been run.

The circular dependencies here make arriving at perfect estimates at any given stage challenging. In the remainder of this chapter, we will present a heuristic algorithm that allows for good– if not optimal– choices of Δx, Δt, and t_0 by exploiting an error model estimated through a low-cost exploration phase.

4.4 *Explore-exploit algorihm*

Consider now the simulation of $J_{T,hp}$ without prior knowledge of the error behavior, subject to a fixed total budget $\mathcal{B}$. We can choose to subdivide the total budget into any number of simulations, such that

$$\mathcal{B} = \sum_n \mathcal{C}_t^{(n)}, \tag{4.6}$$

where $\mathcal{C}_t^{(n)}$ gives the cost of the n-th simulation.

If we knew the error behavior a priori, the optimal solution would be simple: solve for $J_{T,hp}$ at the optimal discretization with one simulation at $\mathcal{C}_t = \mathcal{B}$. In lieu of this knowledge, we take an explore-exploit strategy. Thus we divide the budget into some number of exploration simulations and a final exploit simulation:

$$\mathcal{B} = \underbrace{\sum_m \mathcal{C}_t^{(m)}}_{\mathcal{B}^{\text{explore}}} + \underbrace{\mathcal{C}_t^{\text{exploit}}}_{\mathcal{B}^{\text{exploit}}} \tag{4.7}$$

Using the explore budget, $\mathcal{B}^{\text{explore}}$, we will run M explore simulations, whose goal is to determine estimates of Δx_{opt}, Δt_{opt}, $T_{s,\text{opt}}$, and $t_{0,\text{opt}}$ that should minimize the error in the exploit stage, rather than making an accurate estimate of J_∞ on their own. Given the resulting estimate, we then use the remaining $\mathcal{B}^{\text{exploit}}$ to compute $J_{T,hp}$ with the goal of making $e_{T,hp}$ as small as possible, using the best guess given the knowledge accumulated in the exploration stage.

Transient-robust small-sample simulation

In Algorithm 1, we show a proposed heuristic explore-exploit procedure. We denote by Θ^* the Bayesian random variable (RV) description of the model θ^*, which which is functionally generated by the Hamiltonian Monte Carlo sampler detailed in Section 2.2; realizations of this RV are denoted by θ^*. Additionally, we denote by $(\cdot)^+$ a quantity that is based on the *observed* transient behavior, rather than the *expected* behavior, denoted by $\tilde{(\cdot)}$. For the expected transient behvior, we use a shorthand max function to denote a conservative estimate for the transient behavior, such that $\tilde{\psi} = (\max_n |A_\lambda^{(n)}|, \max_n T_\lambda^{(n)})$. Last but not least,

$$e_{\text{model}}(\Delta x, \Delta t, t_0) = e_{\text{model}}(\Delta x, \Delta t, t_0 | \theta(\theta^*), \psi(\psi^*))$$

is given by the form in (3.27).

4.5 Numerical results: algorithm application

In this section, we demonstrate the use of Algorithm 1. We will run a set of 20 RK3 discretizations of the KSE-ODE problem with $N_t = 4 \times 10^4$ for exploration, and assess the error performance that they might project to have on a simulation with $N_t = 3.2 \times 10^6$, i.e. using $\mathcal{B}\Delta x / \mathcal{C}_{p_x p_t} = 4 \times 10^6$. As previously, we will reject simulations that lack the fidelity to be in the convergent regime of Δt and T_s given by Table 3.1. In addition to the convergence limit on T_s, we also will require $T > 1200$ in order to guarantee that the transient behavior is satisfactorially captured. In order to choose the next stage's Δt, and t_0, we take a very naïve approach, sampling a set $\{\theta\} \sim \Theta^*$, calculating Δt_{opt} and $t_{0,\text{opt}}$ with each sample, then choosing randomly among the implied sets $\{(\Delta t_{\text{opt}}, t_{0,\text{opt}})\}$ that fall within the constraints. If no member of the set $\{(\Delta t_{\text{opt}}, t_{0,\text{opt}})\}$ falls within the constraints, Δt is chosen as a loguniform sample between the minimum and maximum Δt possible under the constraints.

For the RK3 KSE-ODE case, cost reduces to some constant times the total number of timesteps, N_t.

For the KSE-ODE problem there is no Δx, but this naïve approach can be extended to a full PDE problem with the treatment of Δx and some choices about the sampling process in Δx, Δt, and T_s.

In Figures 4.9 and 4.10 we show the sequences of $\Delta t^{(m)}$ and $T^{(m)}$ and $A_\lambda^{*,(m)}$ and $T_\lambda^{*,(m)}$, as they are generated by the exploration cycle. These values, once computed, cannot be changed without running a new simulation.

In order to quantify the progression of the scheme, we will compare to the optimal values for the exploit simulation with $N_t = 3.2 \times 10^6$ under the model described with $t_0 = 3000$ in Table 4.4. At these values, and using the 95th percentile transient characteristics, we should simulate at the reference optima, given by:

$$\begin{aligned}(\Delta t_{\text{opt}})_{\text{ref}} &= 0.0510 \\ (t_{0,\text{opt}})_{\text{ref}} &= 7.75 \times 10^3 \\ (T_{s,\text{opt}})_{\text{ref}} &= 1.55 \times 10^5 . \\ (e_{\text{opt}})_{\text{ref}} &= 0.276\end{aligned} \tag{4.8}$$

Algorithm 1: Heuristic cycle algorithm for identification of θ and ψ

```
create cost schedule $\mathcal{C}_t^{(m)}, \mathcal{C}_t^{\text{exploit}}$ s.t. $\sum_{m=1}^{M} \mathcal{C}_t^{(m)} + \mathcal{C}_t^{\text{exploit}} = \mathcal{B}$    ▷ many approaches possible
$\{J_{T,hp}\} \leftarrow \varnothing$
for $m = 1, \dots, M$ do
    $\Theta^* \leftarrow \text{HMC}(\theta^* \mid \{J_{T,hp}\})$    ▷ random variable
    if $m = 1$ then
        $\tilde{\psi}^{(m)} = (T_{\text{guess}}, 0)$
    else
        $\tilde{\psi}^{(m)} \leftarrow \max\{\psi(\psi^{*,(n)})\}_{n=1}^{m-1}$
    end if
    choose $(\Delta x^{(m)}, \Delta t^{(m)}, t_0^{(m)})$ using $\Theta^*$ and $\tilde{\psi}^{(m)}$    ▷ many approaches possible
    $N_t^{(m)} \leftarrow \mathcal{C}_t^{(m)} \Delta x^{(m)} / (\mathcal{C}_{p_x p_t} L)$
    $T_s^{(m)} \leftarrow N_t^{(m)} \Delta t^{(m)} - t_0^{(m)}$
    $\mathbf{u}_{hp}^{(m)}(t) \leftarrow \text{PDE}_{hp}(\Delta x^{(m)}, \Delta t^{(m)}, N_t^{(m)})$
    $\psi^{*,(m)} \leftarrow \text{decay}(g_{hp}(\mathbf{u}_{hp}^{(m)}))$
    $t_0^{+,(m)} \leftarrow \text{argmin}_{t_0}(e_{\text{model}}(t_0, \Delta x^{(m)}, \Delta t^{(m)}))$    ▷ $\Delta x^{(m)}$ and $\Delta t^{(m)}$ fixed.
    $T_s^{+,(m)} \leftarrow N_t^{(m)} - t_0^{+,(n)}$
    $J_{T,hp}^{(m)} \leftarrow \int_{t_0^{+,(m)}}^{t_0^{+,(m)} + T_s^{+,(m)}} g_{hp}(\mathbf{u}_{hp}^{(m)}) \, \mathrm{d}t$    ▷ reintegrate optimally
    for $n = 1, \dots, m-1$ do
        $t_0^{+,(n)} = \text{argmin}_{t_0}(e_{\text{model}}(t_0, \Delta x^{(n)}, \Delta t^{(n)}))$    ▷ $\Delta x^{(n)}$ and $\Delta t^{(n)}$ fixed.
        $T_s^{+,(n)} = N_t^{(m)} - t_0^{+,(n)}$
        $J_{T,hp}^{(n)} \leftarrow \int_{t_0^{(n)}}^{t_0^{(n)} + T_s^{(n)}} g_{hp}(\mathbf{u}_{hp}^{(n)}) \, \mathrm{d}t$    ▷ reintegrate optimally
    end for
    $\{J_{T,hp}\} \leftarrow \{J_{T,hp}^{(n)}\}_{n=1}^{m}$    ▷ update & append
end for
$\theta^*_{\text{MAP}} \leftarrow \text{M.A.P.}(\text{HMC}(\theta^* \mid \{J_{T,hp}\}))$
$\tilde{\psi} \leftarrow \max\{\psi(\psi^{*,(n)})\}_{n=1}^{m}$
$(\Delta x^{\text{exploit}}, \Delta t^{\text{exploit}}, t_0^{\text{exploit}}) \leftarrow \Big((\Delta x_{\text{opt}}, \Delta t_{\text{opt}}, t_{0,\text{opt}}) \;\Big|\; \Big(\theta(\theta^*_{\text{MAP}}), \tilde{\psi}, \mathcal{C}_t^{\text{exploit}}\Big)\Big)$
$N_t^{\text{exploit}} \leftarrow \mathcal{C}_t^{\text{exploit}} \Delta x^{\text{exploit}} / (\mathcal{C}_{p_x p_t} L)$
$T_s^{\text{exploit}} \leftarrow N_t^{\text{exploit}} \Delta t^{\text{exploit}}$
$\mathbf{u}_{hp}^{\text{exploit}}(t) \leftarrow \text{PDE}_{hp}(\Delta x^{\text{exploit}}, \Delta t^{\text{exploit}}, N_t^{\text{exploit}})$    ▷ final state estimate
$\psi^{*,(m)} \leftarrow \text{decay}(g_{hp}(\mathbf{u}_{hp}^{\text{exploit}}))$
$J_{T,hp}^{\text{exploit}} \leftarrow \int_{t_0^{\text{exploit}}}^{t_0^{\text{exploit}} + T_s^{(n)}} g_{hp}(\mathbf{u}_{hp}^{\text{exploit}}) \, \mathrm{d}t$    ▷ final output estimate
$\theta^* \leftarrow \text{M.A.P.}(\text{HMC}(\theta^* \mid \{\{J_{T,hp}\}, J_{T,hp}^{\text{exploit}}\}))$    ▷ allows final output error estimate
```

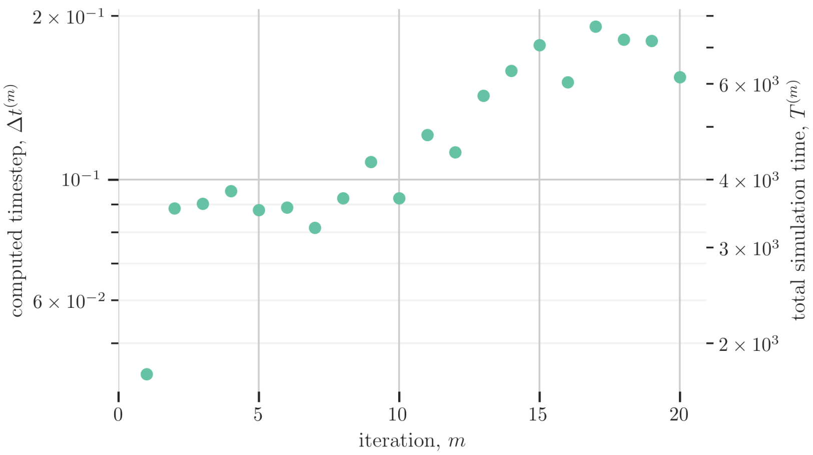


Figure 4.9: Sequence of $\Delta t^{(m)}$ and $T^{(m)}$ values for $M = 20$ explore simulations at $N_t = 4 \times 10^4$ generated according to Algorithm 1.

In Figures 4.11, Figures 4.12, and Figures 4.13, we show the estimates of Δt_{opt}, $t_{0,\text{opt}}$, and $T_{s,\text{opt}}$ for the target $N_t = 3.2 \times 10^6$ simulation as estimated up to the m-th simulation. These show progress towards the optimizer, which appears to be approaching an asymptote with some deviation from– but close to– the reference value.

The most important assessment of the performance of the scheme is how well the exploit-stage simulation might be expected to perform at a given exploration stage. We can do this using the error under the model in (3.27) given the reference values, evaluated at the estimated optimizer:

$$e_{\text{ref}}(t_{0,\text{opt}}, \Delta x_{\text{opt}}, \Delta t_{\text{opt}}) = e_{\text{model}}(t_{0,\text{opt}}, \Delta x_{\text{opt}}, \Delta t_{\text{opt}} \mid \theta_{\text{ref}}, \psi_{\text{ref}})$$

This is done in Figure 4.14. As we can see from the results, after a spontaneously good initial point from the prior, Algorithm 1 manages to control the error a final value around 1.5 times the optimal in expectation, which would occur if the exploit stage timesteps $N_t = 3.2 \times 10^6$ were used with perfect a priori knowledge. At $m = 2$ and $m = 3$, for example, the iterative process would have arrived at errors a factor of seven higher than the optimal.

After the final stage, we make a final posterior estimate to optimize the exploit stage, resulting in:

$$\begin{aligned} \Delta t_{\text{exploit}} &= 0.0394 \\ t_{0,\text{exploit}} &= 8.44 \times 10^3 \\ T_{s,\text{exploit}} &= 1.18 \times 10^5 . \\ e_{\text{opt,exploit}} &= 0.390 \end{aligned} \tag{4.9}$$

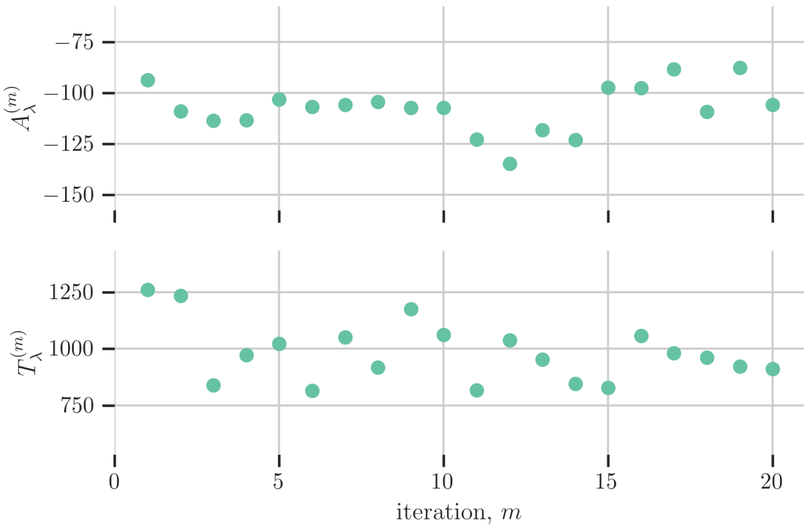


Figure 4.10: Sequence of $\psi^{*,(m)}$ values for $M = 20$ explore simulations at $N_t = 4 \times 10^4$ generated according to Algorithm 1.

Using the reference model, the error at this posterior optimizer is expected to be

$$e_{\text{ref}}(\Delta t_{\text{exploit}}, t_{0,\text{exploit}}, T_{s,\text{exploit}}) = 0.357 \tag{4.10}$$

Thus, though some errors certainly remain, the explore cycle in Algorithm 1 results in good approximants of the reference values in (4.8) which are computed exclusively using the $M = 20$ exploration computations.

It is noted that in the absence of such a result, one must make an arbitrary guess where to simulate the system. One such naïve approach might have been to perform the simulation at the relevant convergence limit in Table 3.1, under the assumption that the discretization errors will be dominated by statistical ones. In this event, the error in the exploit simulation would be at least 5.1, accounting for discretization error alone, and thus the approach herein reduces the error by a factor of at least twelve at $N_t = 3.2 \times 10^6$. While this is only one of many possible approaches, it highlights that the error reduction under this strategy– compared to naïve approaches– is likely to be significant.

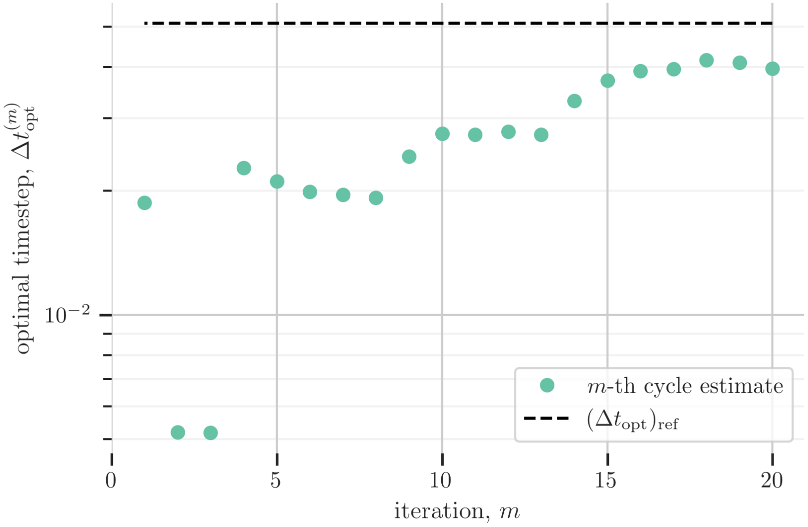


Figure 4.11: Sequence of Δt_{opt} estimates for exploit simulation generated by each of $M = 20$ exploration simulations at $N_t = 4 \times 10^4$.

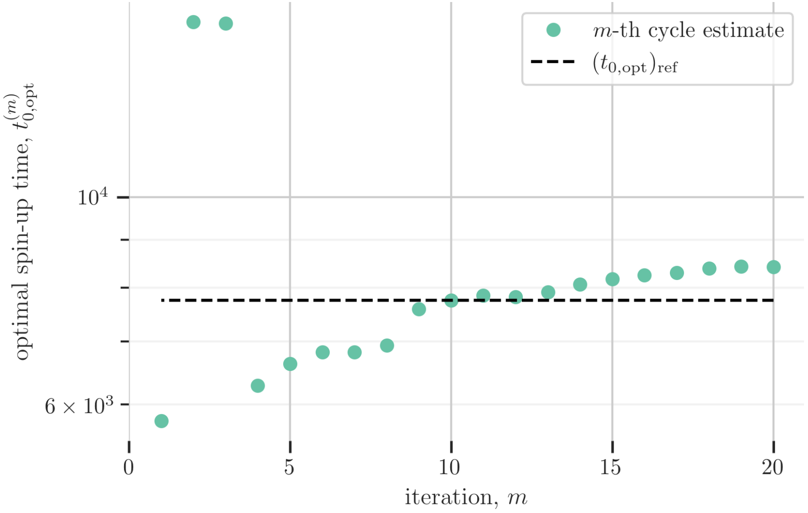


Figure 4.12: Sequence of $t_{0,\mathrm{opt}}$ estimates for exploit simulation generated by each of $M = 20$ exploration simulations at $N_t = 4 \times 10^4$.

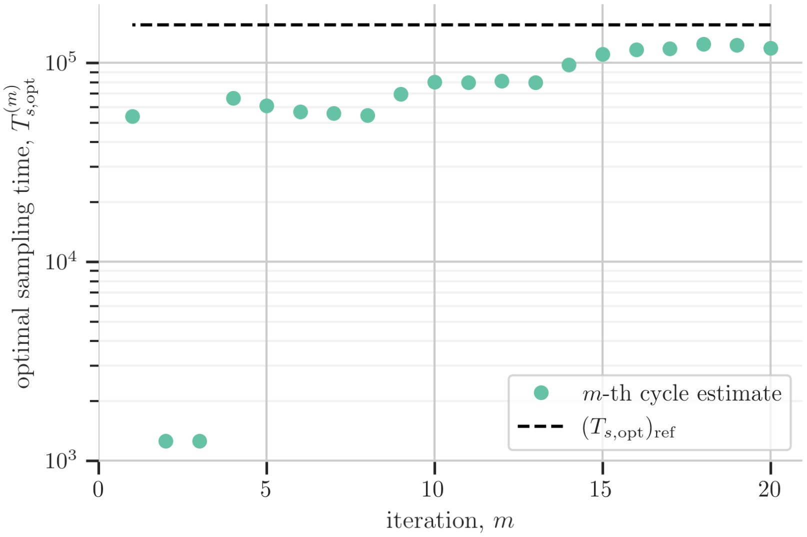


Figure 4.13: Sequence of $T_{s,\text{opt}}$ estimates for exploit simulation generated by each of $M = 20$ exploration simulations at $N_t = 4 \times 10^4$.

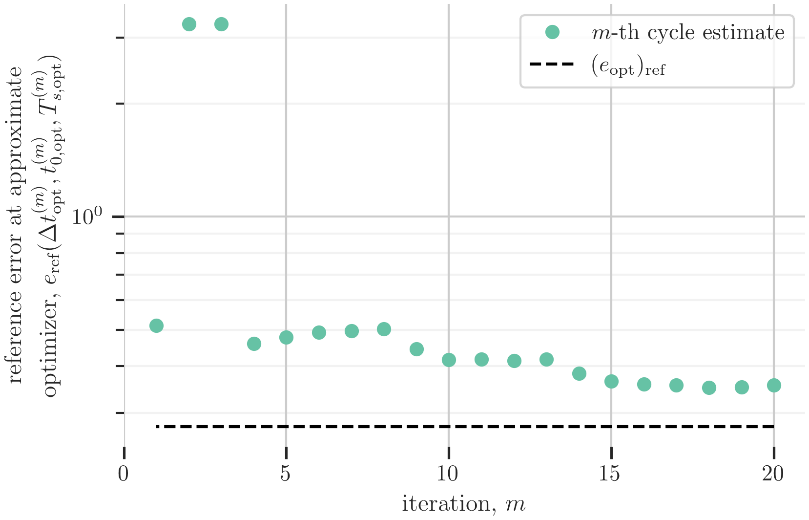


Figure 4.14: Comparison between reference error at best estimate Δt, t_0 for exploit simulation and reference e_{opt} generated by each of $M = 20$ exploration simulations at $N_t = 4 \times 10^4$.

To complete the section, we now run an ensemble of 25 exploit simulations at the optimizer in (4.9), in order to quantify the spread of outcomes possible from the exploit stage. In Figure 4.15, we show the histogram of $J_{T,hp}^{\text{exploit}}$ from the resulting simulations. As we can see by comparing the estimate $e_{\text{opt}} = 0.370$ in (4.9), the algorithm allows a good prediction of the spread of the errors in $J_{T,hp}^{\text{exploit}}$.

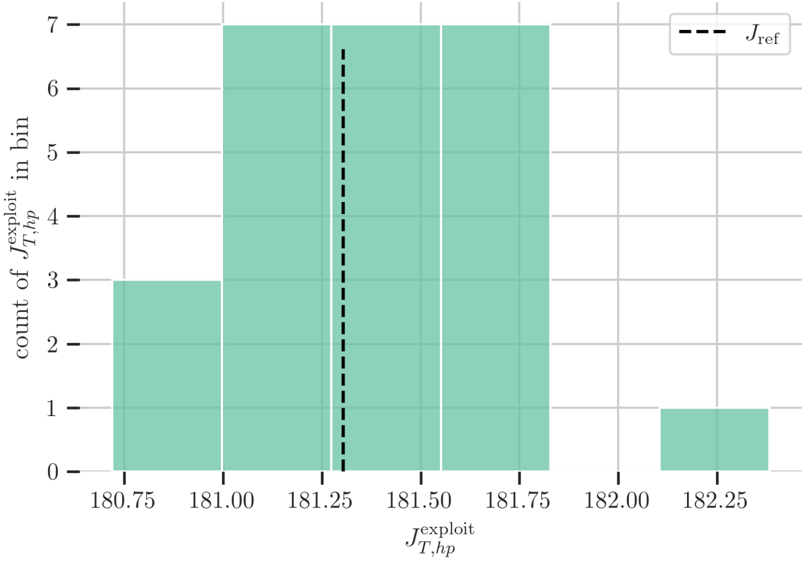


Figure 4.15: Comparison between reference error at best estimate Δt, t_0 for exploit simulation and reference e_{opt} generated by each of $M = 20$ exploration simulations at $N_t = 4 \times 10^4$.

This error estimate can be further refined by incorporating $J_{T,hp}^{\text{exploit}}$ into a final small-sample model estimate, by finding the MAP estimate θ^*_{MAP} incorporating the exploit stage result into the data:

$$\left\{ J_{T,hp}^{(1)}, \ldots, J_{T,hp}^{(M)}, J_{T,hp}^{\text{exploit}} \right\},$$

then calculating e_{model} using ψ^*_{exploit} and the resulting θ^*_{MAP}. In Figure 4.16, we show a histogram of the final MAP error estimate from each $J_{T,hp}^{\text{exploit}}$. As with the accumulated exploration result, we find that the final posterior estimate after the exploit simulation is also descriptive of the spread of errors we see about J_{ref}. This demonstrates that the procedure in Algorithm 1 not only enables near-optimal discretization but also allows for a reliable estimate of $e_{T,hp}$ at the conclusion of the exploit simulation.

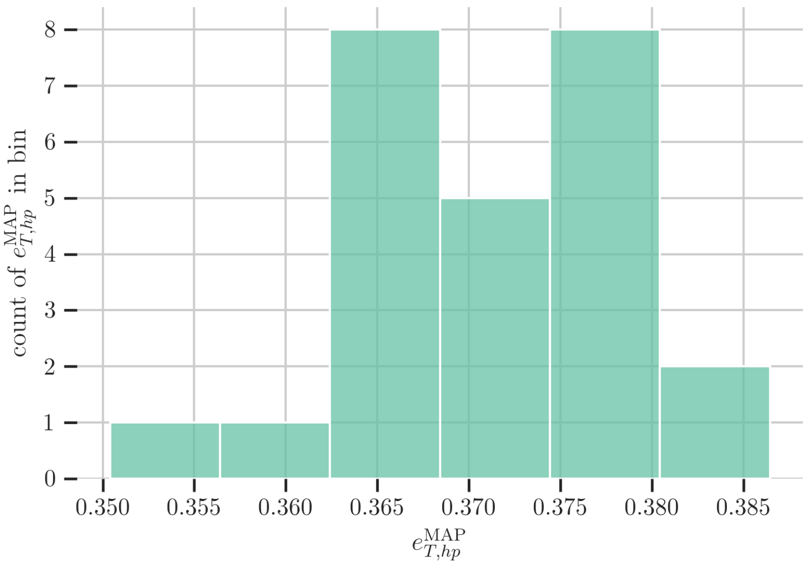


Figure 4.16: Distribution of final estimates at Δt_{opt}, $t_{0,\mathrm{opt}}$, $T_{s,\mathrm{opt}}$ resulting from final MAP estimates.

4.6 *Numerical results: performance expectation*

Finally, in this section, we seek to demonstrate that this process is repeatable. In order to do so, we will repeat the explore stage 100 times and show that the resulting set of optimizers should consistently achieve a low error according to the reference error model.

For each instance of the explore stage, we independently run $M = 20$ explore simulations according to Algorithm 1 and evaluate $\Delta t^{\mathrm{exploit}}$, t_0^{exploit}, and T_s^{exploit} for each. In Figure 4.17, the resulting optimizers of this study are shown on histograms. These results demonstrate that the process can reliably– though not without a few outliers– find a near-optimal discretization for the exploit stage.

We can project the error at these optimizers using the reference values of the error model. In Figure 4.18, the error under e_{ref} implied by Table 4.4 is evaluated at each instance's optimizer to give an estimate of the error at the approximated optimizer. These results show that, of 100 simulations run using the methodology outlined in this chapter, only three are expected to have more than a factor of two more than optimal amount of error, indicating reliable estimation of the optimizer.

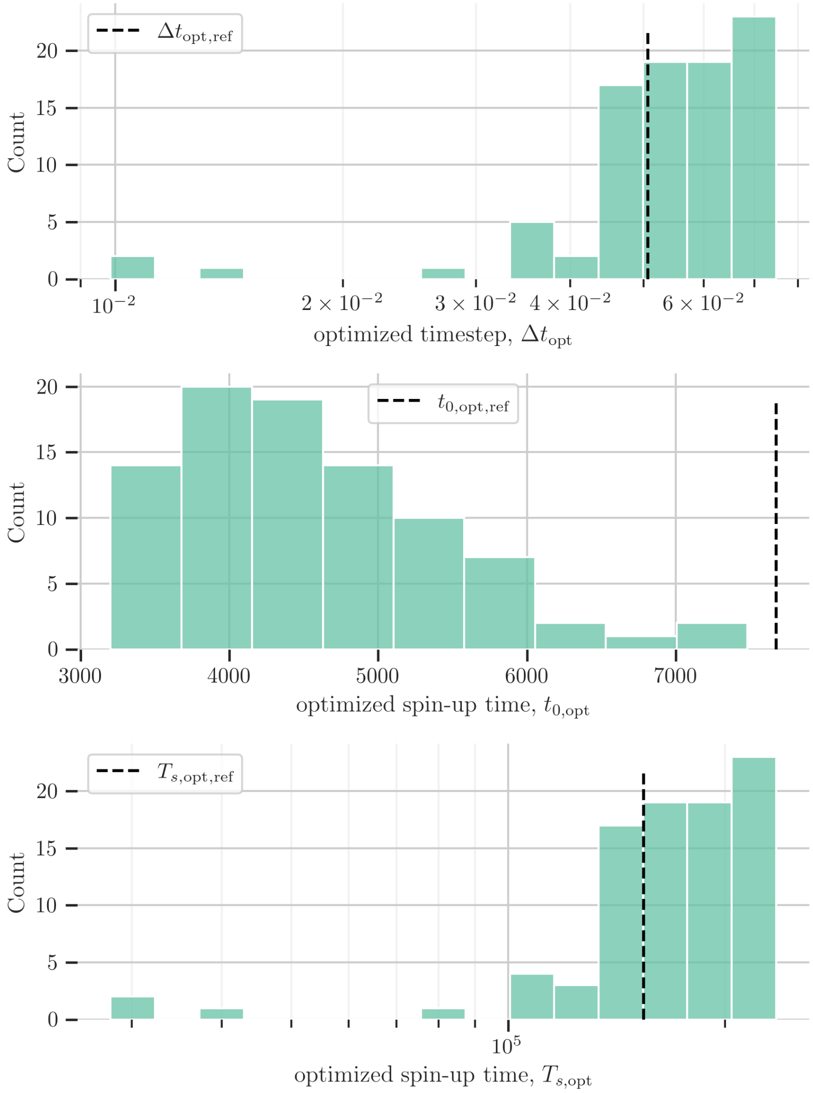


Figure 4.17: Histogram of optimal discretization estimates for exploit simulation after 100 independent runs of the exploration algorithm.

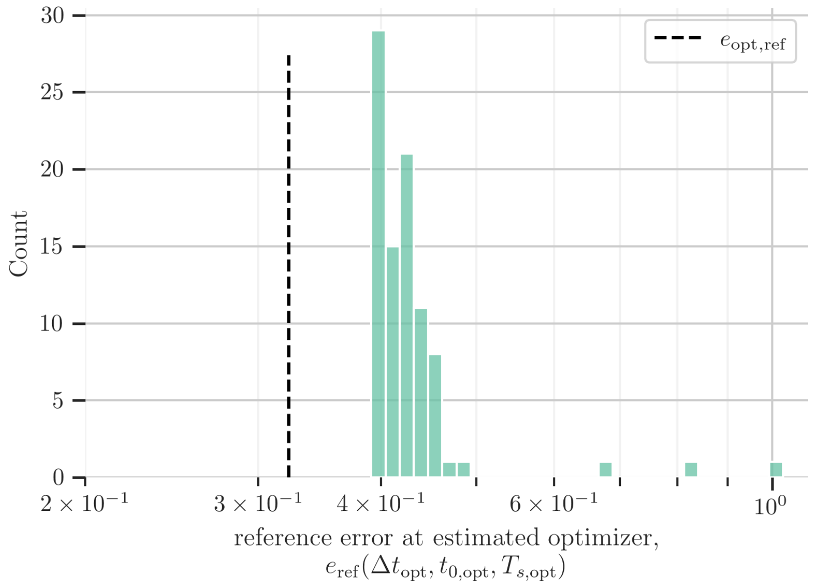


Figure 4.18: Histogram of reference error at approximated optimizers for exploit simulation after 100 independent runs of the exploration algorithm.

4.7 *Conclusion*

In this chapter, we have demonstrated that it is possible to make estimates of the model for the behavior of solutions on the attractor using a data from simulations with non-negligible spin-up transient errors via a correction term. Then, we use this result to demonstrate that by the use of an explore-exploit scheme it is possible to achieve near-optimal results in terms of the balance of fixed-cost error contributions from discretization, spin-up transient, and sampling errors.

There remains significant room for improvement to these results. One key issue is that the deterministic correction approach for the transient component of the error, while adequate, clearly induces some errors that diminish the performance of the small-sample method. Additionally, the framework developed here neglects the correlation between the spatial resolution and the spin-up transient behavior, namely T_λ, which grows as the number of spatial degrees of freedom are increased for the Kuramoto-Sivashinsky equation, though should be expected to converge eventually towards an asymptote. The work of Fernandez and Wang [2017] in particular covers the discretization dependence of the temporal properties– particularly the Lyapunov exponents– of dynamical systems in CFD.

While there remains room for improvement, the novel capability of this method to achieve efficient use of resources has immediate utility in numerics, as well as providing a framework in which a number of key and outstanding questions in computation might be considered. We will look at some of these key questions in the concluding chapter. In terms of the immediate utility, the heuristic algorithm herein gives a foundation upon which many results in optimization can be brought to bear. One particular example is the choice of discretizations in the exploration stages, which is ripe for optimization. A particular approach of interest is an *information-maximizing* experimental design approach[1], in which information theory could be used to make a choice that is optimal in terms of shrinking the uncertainty about control parameters– like $t_{0,\mathrm{opt}}$, Δx_{opt}, and Δt_{opt}– as a function of the error model.

In addition to optimal choices in the exploration stage, another area that is primed for optimization is the allocation of resources for exploration vs. exploitation. The explore-exploit problem has a rich history in multiple academic disciplines, including psychology[2], autonomy[3], and artificial intelligence[4]. In Algorithm 1, we arbitrarily chose to use $M = 20$ simulations for exploration, each 1% of the total simulation budget. It is very likely that more effective and adaptive allocation of costs into exploration and exploitation is possible, especially in the context of the Bayesian method used herein, and this represents another key area for extension of this work.

[1] Xun Huan and Youssef M. Marzouk. Simulation-based optimal Bayesian experimental design for nonlinear systems. *Journal of Computational Physics*, 232(1):288–317, 2013

[2] Daniel J. Navarro, Ben R. Newell, and Christin Schulze. Learning and choosing in an uncertain world: An investigation of the explore-exploit dilemma in static and dynamic environments. *Cognitive Psychology*, 85:43–77, 2016

[3] Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1):99–134, 1998

[4] Kaelbling et al., 1998; Isaac J Sledge and José C Príncipe. Balancing exploration and exploitation in reinforcement learning using a value of information criterion. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 2816–2820. IEEE, 2017; and Jonathan Sorg, Satinder Singh, and Richard L. Lewis. Variance-based rewards for approximate Bayesian reinforcement learning. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, UAI'10, page 564–571, Arlington, Virginia, USA, 2010. AUAI Press

Summary and future work

You certainly usually find something, if you look, but it is not always quite the something you were after.

–J. R. R. Tolkien, The Hobbit

The concentration of this work has been quantifying and minimizing the gap between the modeling equations used for a simulation and the discrete simulation results that result from computational approximation of those equations. For most non-chaotic systems, this description can be fairly simply represented by the convergence of the discrete solution function to a unique exact solution function of the modeling equations in some appropriate measure. In the following section, we will delineate the contributions of this work, which extends our understanding of costs and errors to chaotic systems.

Contributions

Error modeling for mean output quantities of chaotic, ergodic differential equations

For chaotic, ergodic systems, we narrowly define the goal of a simulation as estimating the mean, J_∞, of some quantity of interest. The first major contribution of this work, then, is to propose a comprehensive model for the simulation error in estimates of J_∞. In this model, simulation error consists of a combination of sampling error due to chaotic variation on the attractor, transient error due to spin-up, and discretization errors due to temporal and (when relevant) spatial discretization.

The resulting model offers insight into the fundamental efficiency limitations of approximations of J_∞. The central limit theorem can be expected to govern the sampling behavior for a wide class of ergodic problems, and when this is the case it will limit the asymptotic rates of convergence of the error in a finite-time esimate J_T with respect to J_∞ as a function of cost to the rate $-1/2$. Simulations of chaotic systems will often be a finite-time estimate *of a discrete approximation* of the true system. In this case, simulation results $J_{T,hp}$ are subject to both sampling *and* discretization errors; the rates of convergence of the total simulation error error, given for ODEs in (2.4) and (3.19) for PDEs, will not reach the $-1/2$ rate expected under the CLT, but will have a q-dependent rate penalty. These discrete rates approach the CLT rate as $q \rightarrow \infty$, where q is the convergence rate of the discretization error. While significant improvements in the rates disappear in the chaotic, ergodic analysis, our work demonstrates that the use of high-order schemes can nonetheless have significant reductions in error at a given amount of wallclock cost, at least for the systems shown in the work.

In addition, we show that the optimal balance of error contributions can take different forms than the CLT-limited convergence rate. When budget limitations dominate, we show a limiting convergence with T_s^{-1}, associated with the reduction of the spin-up transient error. To the author's knowledge, this work represents a first attempt to quantify an optimal use of resources in this budget-constrained limit.

Last but not least, we incorporate the possibility of ensemble parallelism, and give an assessment of the potential benefits of ensemble estimation. We show that, as with the sampling time, ensemble estimation of a discrete system cannot improve at the ideal rate for a purely stochastic system, but incurs a penalty when both discretization and sampling errors are present. Moreover, we show that in order to reduce the expected error by parallel ensemble estimation, the specified discretization must be continually optimized in order to prevent domination by the discretization error as the number of ensemble instances, M_{ens}, increases. Finally, we account for the fact that the spin-up cost does not amortize across processors, but must be controlled *on every instance* of the simulation, and amend the total cost-error model results to account for this.

Small-sample Bayesian estimation of sampling and discretization errors on the attractors of chaotic, ergodic differential equations

In addition to theoretical description of these errors, the work further demonstrates that a Bayesian inference approach can robustly identify the described error behavior on the attractor using only only a small number of discrete simulations of the system.

While this is not the first use of a Bayesian method in the context of chaotic systems (e.g., Oliver et al. [2014]), a novel contribution of this work is the use of a Bayesian method to identify a model for the errors in a discrete simulation method.

Explore-exploit scheme for near-optimal simulation of chaotic, ergodic differential equations

By amending the small-sample methods to account for the effect of spin-up transient effects, we propose a scheme to explore possible discretizations using some fraction of a budget in order to then exploit the remainder of the budget in a near-optimal manner. Despite multiple naïve heuristic shortcuts and ample room for improvement in the explore-exploit scheme, we demonstrate that near-optimal use of computational resources is possible using the naïve scheme.

Taken as a whole, the error model we propose suggests that current state-of-the-art processes for simulation planning are likely to be subject to cost-error inefficiencies, while our demonstration of the explore-exploit scheme suggests that it is possible to approach the optimal use of resources.

Outstanding issues and improvements

Error modeling in non-convergent regions

One of the fundamental assumptions that is made in this work is that simulations are always performed in the asymptotically convergent regions for the parameters of interest. These have all been identified in this work by exhaustive computation and a priori confinement of the parameter spaces for Δx, Δt, and T_s.

In practice, these will not be known a priori: pre-convergent behavior must be either detected or modeled accurately. One possibility for modeling pre-convergent behavior for the discretization error is by the use of higher-order terms. The discretization error is derived, in effect, by the first non-zero entry in Taylor series approximation of the true solution. It is possible that by incorporating higher terms than the leading order into the modeling process may enable capture of pre-convergent behavior, although this is not guaranteed. For the CLT convergence, short sampling times T_s may also have non-convergent behavior, for which non-normal distributions of the sampling error may be present. While there is not a Taylor series justification for the inclusion of higher-order terms for the CLT, one possible treatment might be to use a variance that scales with T_s^{-1} in the small T_s limit, like the expected transient behavior, but scales with $T_s^{-1/2}$ when T_s becomes large.

A key problem with eother of these approaches is that they have implications for the identifiability of the models, and these implications must be carefully examined before attempting to fit them, especially in the small-sample case. Further investigation of the pre-asymptotic regime is left to future work.

Improved small-sample fitting of data with spin-up transients

An area of this work that needs more attention is the use of transient-corrected data for the small-sample system identification. While the process covered in Chapter 4 achieves its goal, the fits to transient-corrected output data leave much room for improvement.

Fitting of the exponential decay in the outputs of the Kuramoto-Sivashinsky equation, for example, demonstrates that the exponential thesis for the decay behavior has high explanatory value. However, the accuracy of these fits remains unexamined, and the results in Section 4.1 seem to indicate that the resulting transient errors e_λ– used to correct $J_{T,hp}$ against the transient effects– are inaccurate when t_0/T_λ is not sufficiently large.

An alternative scheme might be to treat the "corrected" values in a hierarchical model, which would effectively impute guesses for $J^{\mathcal{A}}_{T,hp}$, the output that would have resulted from the related trajectory on the attractor and to which the simulated trajectory converges as $t \to \infty$. Any such model, however, is likely to struggle to distinguish between the natural variance in the system due to chaos and uncertainty in the transient model, and such a model for this reason must be carefully constructed.

Exploration stage optimization in explore-exploit algorithm

As noted in Chapter 4, the explore-exploit algorithm used therein was based on a number of naïve, heuristic choices in order to develop a proof-of-concept optimization scheme. While the results generated in Ch. 4 show promise, it is very likely that they can be improved.

This area is left to future researchers, but we expect that a combination of canonical approaches for explore-exploit problems and Bayesian experimental design methods are primed to bring significant improvements in accuracy by the conclusion of the exploration stage of the explore-exploit algorithm.

Applications and extensions

Given the results and contributions of this thesis, we now conclude by considering how the framework established in this thesis can be applied and used.

Mixed grid- and ensemble-parallelism for discretization of chaotic PDEs

In this work, we have demonstrated that the CLT imposes a fairly restrictive limit on the relationship between the cost and error of a simulation, and in so doing we have highlighted the importance of advanced methods for speeding up and reducing the error in simulations of chaotic PDEs. Perhaps the foremost candidate among these methods is parallelization. In the body of this thesis, we only considered ensemble parallelism, which we showed could reduce the sampling error in a system but cannot reduce the cost of resolving spin-up.

While our analysis of ensemble parallelism is fairly exhaustive for low-dimensional ODE problems like the Lorenz system, it must be extended to also be able to account for and optimize usage when the high-dimensional spatial system or the temporal advancement scheme can be parallelized within a given instance as well. Suppose P processors are available for a job: the key question is how to allocate processors to M_{ens} ensembles with instances that run with N_p processors each, such that:

$$P = M_{\text{ens}} N_p \tag{c.11}$$

In high-performance computing (HPC), the benefits of parallelization are traditionally described by Amdahl's law[5] for fixed workload com-

[5] Gene M. Amdahl. Validity of the single processor approach to achieving large scale computing capabilities. In *Proceedings of the April 18-20, 1967, Spring Joint Computer Conference*, AFIPS '67 (Spring), page 483–485, New York, NY, USA, 1967. Association for Computing Machinery

puting or Gustafson's model[6] for fixed time computing, with further extensions for for memory-bounded computing[7], heterogeneous computing[8], and general high-performance computing with multiple constraining factors[9]. Amdahl's law for a fixed workload W can be written as:

$$\frac{T_{\text{serial}}}{T_{\text{parallel}}} = \frac{W}{W(1-\rho) + \frac{W\rho}{N_p}} = \frac{1}{(1-\rho) + \frac{\rho}{N_p}}, \tag{c.12}$$

where T_{serial} and T_{parallel} are the computation times of the serial and parallel codes, N_p gives the number of parallel processors, and ρ gives the fraction of the workload W that can be parallelized (with $(1-\rho)$ the portion that cannot be run in parallel).

In (3.45) we develop a model for the wallclock costs of a simulation, which we can rewrite assuming within-instance parallelization benefits are described by Amdahl's law:

$$\mathcal{C}_t^{\text{parallel}} = \left((1-\rho) + \frac{\rho}{N_p}\right) \mathcal{C}_{p_x p_t} N_{\text{elem}} N_t. \tag{c.13}$$

In this case, we assume for simplicity ρ is fixed for a discretization scheme on a given problem. This then propogates into (3.27) where $\mathcal{E}$ can be replaced with

$$\frac{\mathcal{E}}{(1-\rho) + \frac{\rho}{N_p}},$$

allowing, with (c.11):

$$\begin{aligned} e_{\text{model}}^{\text{parallel}} = C_{q_x} \Delta x^{q_x} + B_{q_t} \Delta t^{q_t} + \frac{A_0}{\sqrt{P/N_p}} \left(\frac{\mathcal{E}}{L\left((1-\rho) + \frac{\rho}{N_p}\right)} \Delta x \Delta t - t_0 \right)^{-r} \\ + |A_\lambda| T_\lambda \left(\frac{\mathcal{E}}{L\left((1-\rho) + \frac{\rho}{N_p}\right)} \Delta x \Delta t - t_0 \right)^{-1} \exp\left(-\frac{t_0}{T_\lambda}\right). \end{aligned} \tag{c.14}$$

This leads to a new optimization problem, which would give the optimal use of P processors to minimize $e_{\text{model}}^{\text{parallel}}$. Without solving for the optimizer of (c.14), we can see that if $\rho \approx 1$, i.e. a simulation instance is readily parallelizable, *instance parallelization can reduce the cost of resolving spin-up*, which we have shown in this work ensemble parallelization can not.

Mesh adaptation for chaotic PDEs

Another technique for error reduction is mesh adaptation, which seeks to reduce the error in a simulation by optimizing grids used for

[6] John L. Gustafson. Reevaluating Amdahl's law. *Commun. ACM*, 31(5): 532–533, May 1988

[7] Xian-He Sun and Lionel M Ni. Another view on parallel speedup. In *Supercomputing '90: Proceedings of the 1990 ACM/IEEE Conference on Supercomputing*, pages 324–333, 1990

[8] Mark D. Hill and Michael R. Marty. Amdahl's Law in the multicore era. *Computer*, 41(7):33–38, 2008

[9] Ashur Rafiev, Mohammed A. N. Al-Hayanni, Fei Xia, Rishad Shafik, Alexander Romanovsky, and Alex Yakovlev. Speedup and power scaling models for heterogeneous many-core systems. *IEEE Transactions on Multi-Scale Computing Systems*, 4(3):436–449, 2018

the discretization of a system.

It is far outside the scope of this thesis to suggest a resolution to the question of *how* to achieve grid adaptation for chaotic flows. However, the framework laid out in this thesis can be extended to understand the improvements that might be realized through grid adaptation. The effect of grid adaptation would be, in the terms of (3.13), to replace the discretization error model with an effective error model:

$$e_{hp,x} \approx C_{\mathrm{eff}} \Delta x_{\mathrm{eff}}^{q_{x,\mathrm{eff}}}, \tag{c.15}$$

which is based on some effective measure of characteristic grid length, Δx_{eff}. This effective model would have the benefit of applying to situations where the solution is irregular, for which the simple treatement of discretization error in this work would be expected to come in under the optimal rate (i.e. $q_x < p_x + 1$) but for which the effective model remains optimal $q_x \approx p_x + 1$.

A complete treatment of the extension of the error modeling to anisotropic adapted grids is outside the scope of this work, but the framework herein can be adapted to understand the relationship between adaptive spatial discretizations the total errors. While complete generalization remains outstanding, we remark that the results of this thesis suggest that reduction of discretization error at a given cost must be high in order to overcome the rate limitations that come from the central limit theorem.

Adaptive space-time discretization of chaotic PDEs

In the previous sub-section, our discussion centered around spatial mesh adaptation. Another developing capability is space-time discretization, in which problems are solved as monolithic $(d+1)$-dimensional problems, where d is the number of spatial dimensions. In this context, time is incorporated as if it were an additional spatial dimension, in which temporal flux is interpreted as convection. The result of such a treatment is a monolithic system of equations that can be solved to describe the system at all times in a spatiotemporal domain. One of the key benefits of such an approach is that it enables space-time adaptivity: where spatial adaptation optimizes the grid used from one timestep to the next in a timestepping scheme, a space-time approach allows for computational resources to be allocated preferentially to the regions in space *and time* that are of importance.

Unfortunately the extension of space-time methods to chaotic systems is non-trivial. A key problem is that chaotic sensitivity to initial conditions limits the time window on which the space-time problem can be solved with a Newton-style nonlinear solver, before the linearization becomes ill-conditioned. Suppose a space-time method covers a temporal domain of period ΔT, and it is solved using a linear solver that can tolerate the inversion of matrices with up to some condition number ε. Because the deviation between the solution at t and the solution at $t + \Delta T$ is governed by the Lyapunov stability, we can expect:

$$\mathrm{cond}(\mathbf{J}) \sim \exp(\Lambda_{\mathrm{max}} \Delta T),$$

where $\mathbf{J}$ is the Jacobian of the global problem and Λ_{max} is the largest Lyapunov exponent of the system, which will be positive for a chaotic system. This leads to an expression that should give a limit on the length of time that an individual space-time discretization of a chaotic system can span:

$$\Delta T_{\mathrm{max}} \leq \frac{1}{\Lambda_{\mathrm{max}}} \log \varepsilon, \tag{c.16}$$

under the simple assumptions here.

For chaotic problems, we will need T_s large enough to have a meaningful statistical sample; with some certainty, $T_s \gg \Delta T_{\mathrm{max}}$ when this is the case. In order to overcome this fundamental limitation, a time-slab approach is the most obvious candidate, in which the total simulation time T is subdivided into "time-slabs" of length ΔT, with the final state of any given time-slab used as the initial condition for the next.

Such an approach could have a few significant benefits. For one, it is likely that an improvement in the per-instance parallel speedup would result, as the non-parallelizable portion of a computation will be amortized more favorably across a significant region in time, resulting in $\rho \to 1$ (this is demonstrated for flows in porous media by Jayasinghe [2018]). Moreover, many applications have complex flow in fairly compact space-time regions. For these types of problems, it is possible that the cost benefits from space-time adaptivity may significantly change the cost-error relationship depite the central limit theorem. As with spatial adaptation, the error models in this work need to be revised for these space-time adaptive problems; however the fundamental framework in this work is still valuable as a starting point to understand and quantify the relationship between costs and errors when applying these types of space-time adaptive discretizations.

Epistemological errors

At the highest level, we can think of the goal of simulation as trying to close the gap between the relevant performance of a real system and computational estimates of that performance. This gap breaks down into two parts: the epistemological error that separates perfect answers of the equations used to model the system from the behavior of the real system, and the simulation error that separates the result of computational estimates of those solutions from the true solutions of the modeling equations. The goal of this work has been to understand, estimate, and minimize the second part.

It is worth distinguishing here between the notion of *epistemological errors* as used in simulation, which describe errors due to modeling physical systems with simplified mathematical models, and *epistemic uncertainties*, which refer generally to errors that can be reduced by repeating an experiment. Paradoxically, in most applications one would interpret epistemological errors as a source of so-called *aleatoric uncertainty*, a mutually exclusive category to epistemic uncertainty. At any rate we are referring to *epistemological errors* in this section.

In very general terms, both of these two types of error are always present. In a perfect simulation paradigm, simulation effort will be expended with some ideal balance between the epistemological errors and the simulation errors: below the limit of epistemological error, increased precision in simulation is wasted, since the real system is no better described by a more precise simulation than the next less precise simulation. When this is the case, there are often more useful allocations of effort and time: further exploring the problem's parametric dependencies (e.g. design changes in an engineering system) at a particular fidelity, quantification of uncertainty, taking an extended break, etc.

In practice, understanding the epistemological errors in simulations at a given fidelity is a grand challenge. Literature that attempts to estimate the limits of epistemological barriers of simulation methods like low-fidelity models, RANS, and LES is sparse, let alone literature that attempts to use these various methods in a resource-optimal way. While estimates for the limiting epistemological errors are elusive, multi-level and multi-fidelity frameworks in which these various methods can be embedded have been developed the literature[10].

An interesting, if tangential, research thrust is into epistemic uncertainty quantification for RANS turbulence models, which primarily concentrates on detailed analysis of errors rather than quantification of high-level output errors (see, e.g., Gorlé and Iaccarino [2013]).

We observe that coupling a framework like the one in this work to the design process is a very interesting and fruitful area for future research. The goal of such a coupling would help to allow LES and other chaotic system simulations to be executed and integrated in a way that maximizes their ability to uncover new information about the design questions at hand without excess cost or waste. With that said, however, robust quantification of this type of wholistic cost-error balancing will require significant development before it is practical or advisable to use full-stack optimal resource allocation for design. Nonetheless, the work in this thesis lays out an important and necessary piece for such a scheme to optimally incorporate data

[10] Gianluca Geraci, Michael S. Eldred, and Gianluca Iaccarino. A multifidelity multilevel Monte Carlo method for uncertainty propagation in aerospace applications. In *19th AIAA Non-Deterministic Approaches Conference*, 2017; and Benjamin Peherstorfer, Philip S. Beran, and Karen E. Willcox. Multifidelity Monte Carlo estimation for large-scale uncertainty propagation. In *AIAA Non-Deterministic Approaches Conference*, 2018

from chaotic simulations.

A

Generalized PDE sampling error model

In this appendix, we will make a general error model for d-dimensional DG-like discretizations under the error model in Chapter 3. The primary distinction between the DIRK-DAE/DGBR4 discretization used herein and discretizations of turbulent Navier-Stokes is that the cost model in Chapter 3 levers savings due to the tri-diagonal linear systems that exist in the 1D case, and these savings do not exist in the d-dimensional case.

We will assume a generic d-dimensional discontinuous Galerkin discretization with order p_x polynomial representations on simplex elements. In this case, each element will have

$$N_{\mathrm{DOF}}^{\mathrm{elem}} = \frac{1}{d!}\frac{(p_x + d)!}{p_x!} \tag{A.1}$$

degrees of freedom. This allows us to write the total number of degrees of freedom in the spatial system as:

$$N_{\mathrm{DOF}}^{\mathrm{sys},x} = \frac{1}{d!}\frac{(p_x + d)!}{p_x!} N_{\mathrm{elem}}. \tag{A.2}$$

Now, we will need to solve such as system with a nonlinear solver. We will assume that the nonlinear solves require number of solutions of the linearized system, and that that number has a well-defined average, $\bar{N}_{\mathrm{NLI}}$. At each iteration, the linearized system with $N_{\mathrm{DOF}}^{\mathrm{sys},x}$ degrees of freedom will be solved. A general statement for the cost of solving a linear system is that it will scale as:

$$\mathcal{C}_{\mathrm{linear}} = C_{\mathrm{LS}}(N_{\mathrm{DOF}}^{\mathrm{sys},x})^{\xi}, \tag{A.3}$$

where ξ is a constant, and C_{LS} is a constant coefficient. For tridiagonal linear systems, as in the 1D KSE DGBR4 case, $\xi = 1$. Most generally, $\xi = 3$ is the upper limit, at which the Gaussian elimination algorithm can be used to solve any linear system. In general, smaller values of ξ might be possible using more advanced or tailored algorithms. Taking these together, the cost of solving the spatial system is:

The block Thomas algorithm for the tridiagonal system ($\xi = 1$) is just one particular example of a specialized value.

$$\mathcal{C}_{\text{sys},x} = \underbrace{\bar{N}_{\text{NLI}} C_{\text{LS}} \left(\frac{1}{d!} \frac{(p_x + d)!}{p_x!} \right)^{\xi}}_{C_{\text{sys},x}} N_{\text{elem}}^{\xi}, \tag{A.4}$$

where $C_{\text{sys},x} = C_{\text{sys},x}(d, p_x)$ is constant for a given physical system and discretization.

The physical system is now solved in the temporal discretization. We will assume an implicit method-of-lines discretization with s stages, at each of which the spatial system must be solved. Thus, the total cost should scale as:

$$\begin{aligned} \mathcal{C} &= s N_t \mathcal{C}_{\text{sys},x} \\ &= \underbrace{s \bar{N}_{\text{NLI}} C_{\text{LS}} \left(\frac{1}{d!} \frac{(p_x + d)!}{p_x!} \right)^{\xi}}_{C_{\text{cost}}} N_{\text{elem}}^{\xi} N_t \end{aligned} \tag{A.5}$$

over N_t timesteps, where $C_{\text{cost}} = C_{\text{cost}}(d, p_x, s)$ is the total cost coefficient, constant for a given physical system and discretization.

Now, we note that $N_{\text{elem}} \approx (L/\Delta x)^d$ and $N_s = T_s/\Delta t$, so that we can write the cost model in terms of Δx, Δt, and T_s:

$$\mathcal{C}_s = C_{\text{cost}} \frac{L^{\xi d} T_s}{\Delta x^{\xi d} \Delta t}, \tag{A.6}$$

which inverts to give T_s:

$$T_s = \frac{\mathcal{C}_s}{C_{\text{cost}} L^{\xi d}} \Delta x^{\xi d} \Delta t. \tag{A.7}$$

Now, we can insert this value into (3.13). The result takes an minimal error e_{opt} at:

$$\begin{aligned} \Delta x_{\text{opt}} = {} & L^{\frac{q_t r \xi d}{q_x q_t + q_x r + q_t r \xi d}} q_x^{-\frac{q_t + r}{q_x q_t + q_x r + q_t r \xi d}} q_t^{\frac{r}{q_x q_t + q_x r + q_t r \xi d}} r^{\frac{q_t}{q_x q_t + q_x r + q_t r \xi d}} (\xi d)^{\frac{q_t + r}{q_x q_t + q_x r + q_t r \xi d}} \\ & C_{p_x}^{-\frac{q_t + r}{q_x q_t + q_x r + q_t r \xi d}} B_{p_t}^{\frac{r}{q_x q_t + q_x r + q_t r \xi d}} A_0^{\frac{q_t}{q_x q_t + q_x r + q_t r \xi d}} M_{\text{ens}}^{-\frac{q_t}{2(q_x q_t + q_x r + q_t r \xi d)}} (\mathcal{C}_s / C_{\text{cost}})^{-\frac{q_t r}{q_x q_t + q_x r + q_t r \xi d}} \end{aligned} \tag{A.8}$$

$$\begin{aligned} \Delta t_{\text{opt}} = {} & L^{\frac{q_x r \xi d}{q_x q_t + q_x r + q_t r \xi d}} q_x^{\frac{r \xi d}{q_x q_t + q_x r + q_t r \xi d}} q_t^{-\frac{q_x + r \xi d}{q_x q_t + q_x r + q_t r \xi d}} r^{\frac{q_x}{q_x q_t + q_x r + q_t r \xi d}} (\xi d)^{-\frac{r \xi d}{q_x q_t + q_x r + q_t r \xi d}} \\ & C_{p_x}^{\frac{r \xi d}{q_x q_t + q_x r + q_t r \xi d}} B_{p_t}^{-\frac{q_x + r \xi d}{q_x q_t + q_x r + q_t r \xi d}} A_0^{\frac{q_x}{q_x q_t + q_x r + q_t r \xi d}} M_{\text{ens}}^{-\frac{q_x}{2(q_x q_t + q_x r + q_t r \xi d)}} (\mathcal{C}_s / C_{\text{cost}})^{-\frac{q_x r}{q_x q_t + q_x r + q_t r \xi d}} \end{aligned} \tag{A.9}$$

$$\begin{aligned} T_{s,\mathrm{opt}} = {} & L^{-\frac{q_x q_t \xi d}{q_x q_t + q_x r + q_t r \xi d}} q_x^{-\frac{q_t \xi d}{q_x q_t + q_x r + q_t r \xi d}} q_t^{-\frac{q_x}{q_x q_t + q_x r + q_t r \xi d}} r^{\frac{q_x + q_t \xi d}{q_x q_t + q_x r + q_t r \xi d}} (\xi d)^{\frac{q_t \xi d}{q_x q_t + q_x r + q_t r \xi d}} \\ & C_{p_x}^{-\frac{q_t \xi d}{q_x q_t + q_x r + q_t r \xi d}} B_{p_t}^{-\frac{q_x}{q_x q_t + q_x r + q_t r \xi d}} A_0^{\frac{q_x + q_t \xi d}{q_x q_t + q_x r + q_t r \xi d}} M_{\mathrm{ens}}^{-\frac{q_x + q_t \xi d}{2(q_x q_t + q_x r + q_t r \xi d)}} (\mathcal{C}_s / C_{\mathrm{cost}})^{\frac{q_x q_t}{q_x q_t + q_x r + q_t r \xi d}} \end{aligned} \tag{A.10}$$

at which:

$$\begin{aligned} e_{\mathrm{opt}} = {} & (q_x q_t + q_x r + q_t r \xi d) L^{\frac{q_x q_t r \xi d}{q_x q_t + q_x r + q_t r \xi d}} q_x^{-\frac{q_x q_t + q_x r}{q_x q_t + q_x r + q_t r \xi d}} q_t^{-\frac{q_x q_t + q_t r \xi d}{q_x q_t + q_x r + q_t r \xi d}} r^{-\frac{q_x r + q_t r \xi d}{q_x q_t + q_x r + q_t r \xi d}} (\xi d)^{-\frac{q_t r \xi d}{q_x q_t + q_x r + q_t r \xi d}} \\ & C_{p_x}^{\frac{q_t r \xi d}{q_x q_t + q_x r + q_t r \xi d}} B_{p_t}^{\frac{q_x r}{q_x q_t + q_x r + q_t r \xi d}} A_0^{\frac{q_x q_t}{q_x q_t + q_x r + q_t r \xi d}} M_{\mathrm{ens}}^{-\frac{q_x q_t}{2(q_x q_t + q_x r + q_t r \xi d)}} (\mathcal{C}_s / C_{\mathrm{cost}})^{-\frac{q_x q_t r}{q_x q_t + q_x r + q_t r \xi d}} . \end{aligned} \tag{A.11}$$

The resulting scaling factors have similar trends to the results in Chapter 3 for the 1D block tridiagonal system. The costs scale at a rate given by:

$$-\frac{q_x q_t r}{q_x q_t + q_x r + q_t r \xi d}.$$

Assuming Gaussian elimination, $\xi = 3$ on a 1D system $d = 1$, and the resulting rates various choices of $q = q_x = q_t$ are shown in Table A.1.

q	1	2	3	4	5	$\cdots$	∞
$-\frac{q_x q_t r}{q_x q_t + q_x r + q_t r \xi d}$	$-1/6$	$-1/4$	$-3/10$	$-1/3$	$-5/14$	$\cdots$	$-1/2$

Table A.1: Sampling cost-error convergence rates for $\xi = 3$ (Dense Gaussian elimination).

Thus, the effect of more computationally complex solves is to slow the convergence towards the CLT rate of $1/2$ in a greater manner than in the tridiagonal case.

Tridiagonal system reduction

For a 1D block tridiagonal system, $\xi \to 1$ and $d = 1$, which results in:

$$\begin{aligned} \Delta x_{\mathrm{opt}} = {} & L^{\frac{q_t r}{q_x q_t + q_x r + q_t r}} q_x^{-\frac{q_t + r}{q_x q_t + q_x r + q_t r}} q_t^{\frac{r}{q_x q_t + q_x r + q_t r}} r^{\frac{q_t}{q_x q_t + q_x r + q_t r}} \\ & C_{p_x}^{-\frac{q_t + r}{q_x q_t + q_x r + q_t r}} B_{p_t}^{\frac{r}{q_x q_t + q_x r + q_t r}} A_0^{\frac{q_t}{q_x q_t + q_x r + q_t r}} \\ & M_{\mathrm{ens}}^{-\frac{q_t}{2(q_x q_t + q_x r + q_t r)}} (\mathcal{C}_s / C_{\mathrm{cost}})^{-\frac{q_t r}{q_x q_t + q_x r + q_t r}} \end{aligned} \tag{A.12}$$

$$\begin{aligned} \Delta t_{\mathrm{opt}} = {} & L^{\frac{q_x r}{q_x q_t + q_x r + q_t r}} q_x^{\frac{r}{q_x q_t + q_x r + q_t r}} q_t^{-\frac{q_x + r}{q_x q_t + q_x r + q_t r}} r^{\frac{q_x}{q_x q_t + q_x r + q_t r}} \\ & C_{p_x}^{\frac{r}{q_x q_t + q_x r + q_t r}} B_{p_t}^{-\frac{q_x + r}{q_x q_t + q_x r + q_t r}} A_0^{\frac{q_x}{q_x q_t + q_x r + q_t r}} \\ & M_{\mathrm{ens}}^{-\frac{q_x}{2(q_x q_t + q_x r + q_t r)}} (\mathcal{C}_s / C_{\mathrm{cost}})^{-\frac{q_x r}{q_x q_t + q_x r + q_t r}} \end{aligned} \tag{A.13}$$

$$\begin{aligned} T_{s,\mathrm{opt}} = {} & L^{-\frac{q_x q_t}{q_x q_t + q_x r + q_t r}} q_x^{-\frac{q_t}{q_x q_t + q_x r + q_t r}} q_t^{-\frac{q_x}{q_x q_t + q_x r + q_t r}} r^{\frac{q_x + q_t}{q_x q_t + q_x r + q_t r}} \\ & C_{p_x}^{-\frac{q_t}{q_x q_t + q_x r + q_t r}} B_{p_t}^{-\frac{q_x}{q_x q_t + q_x r + q_t r}} A_0^{\frac{q_x + q_t}{q_x q_t + q_x r + q_t r}} \\ & M_{\mathrm{ens}}^{-\frac{q_x + q_t}{2(q_x q_t + q_x r + q_t r)}} (\mathcal{C}_s / C_{\mathrm{cost}})^{\frac{q_x q_t}{q_x q_t + q_x r + q_t r}} \end{aligned} \tag{A.14}$$

at which:

$$\begin{aligned}e_{\text{opt}} = (q_x q_t + q_x r + q_t r) L^{\frac{q_x q_t r}{q_x q_t + q_x r + q_t r}} q_x^{-\frac{q_x q_t + q_x r}{q_x q_t + q_x r + q_t r}} q_t^{-\frac{q_x q_t + q_t r}{q_x q_t + q_x r + q_t r}} r^{-\frac{q_x r + q_t r}{q_x q_t + q_x r + q_t r}} \\ C_{p_x}^{\frac{q_t r}{q_x q_t + q_x r + q_t r}} B_{p_t}^{\frac{q_x r}{q_x q_t + q_x r + q_t r}} A_0^{\frac{q_x q_t}{q_x q_t + q_x r + q_t r}} \\ M_{\text{ens}}^{-\frac{q_x q_t}{2(q_x q_t + q_x r + q_t r)}} (\mathcal{C}_s / C_{\text{cost}})^{-\frac{q_x q_t r}{q_x q_t + q_x r + q_t r}} .\end{aligned} \tag{A.15}$$

It is often desirable to use matched convergence rates in the spatial and temporal discretization methods, such that $q_x = q_t = q$:

$$\begin{aligned}\Delta x_{\text{opt}} = L^{\frac{r}{q+2r}} q^{-\frac{1}{q+2r}} r^{\frac{1}{q+2r}} \\ C_{p_x}^{-\frac{q+r}{q^2+2qr}} B_{p_t}^{\frac{r}{q^2+2qr}} A_0^{\frac{1}{q+2r}} \\ M_{\text{ens}}^{-\frac{1}{2(q+2r)}} (\mathcal{C}_s / C_{\text{cost}})^{-\frac{r}{q+2r}}\end{aligned} \tag{A.16}$$

$$\begin{aligned}\Delta t_{\text{opt}} = L^{\frac{r}{q+2r}} q^{-\frac{1}{q+2r}} r^{\frac{1}{q+2r}} \\ C_{p_x}^{\frac{r}{q^2+2qr}} B_{p_t}^{-\frac{q+r}{q^2+2qr}} A_0^{\frac{1}{q+2r}} \\ M_{\text{ens}}^{-\frac{1}{2(q+2r)}} (\mathcal{C}_s / C_{\text{cost}})^{-\frac{r}{q+2r}}\end{aligned} \tag{A.17}$$

$$\begin{aligned}T_{s,\text{opt}} = L^{-\frac{q}{q+2r}} q^{-\frac{2}{q+2r}} r^{\frac{2}{q+2r}} \\ C_{p_x}^{-\frac{1}{q+2r}} B_{p_t}^{-\frac{1}{q+2r}} A_0^{\frac{2}{q+2r}} \\ M_{\text{ens}}^{-\frac{1}{q+2r}} (\mathcal{C}_s / C_{\text{cost}})^{\frac{2}{q+2r}}\end{aligned} \tag{A.18}$$

at which:

$$\begin{aligned}e_{\text{opt}} = (q^2 + 2qr) L^{\frac{qr}{q+2r}} q^{-\frac{2(q+r)}{q+2r}} r^{-\frac{2r}{q+2r}} \\ C_{p_x}^{\frac{r}{q+2r}} B_{p_t}^{\frac{r}{q+2r}} A_0^{\frac{q}{q+2r}} \\ M_{\text{ens}}^{-\frac{q}{2(q+2r)}} (\mathcal{C}_s / C_{\text{cost}})^{-\frac{qr}{q+2r}} .\end{aligned} \tag{A.19}$$

B

Stabilized discontinuous Galerkin/DIRK-DAE method for unsteady fourth-order physical operators

B.1 Background/Introduction

For this thesis, we wanted to find a cheap spatiotemporal system that exhibits chaos as a useful toy problem. Kuramoto-Sivashinsky and its generalizations are the well-established option for this type of problem:

$$\begin{aligned} \frac{\partial u}{\partial t} + \nabla \cdot \left(\boldsymbol{c} u + \frac{\boldsymbol{\alpha}}{2} u^2 \right) + \nabla \cdot \left(\beta \nabla u + \gamma \nabla^3 u \right) &= f && \text{in } \Omega \\ u &= u_d && \text{on } \Gamma_{B'} \\ \nabla u \cdot \hat{n} &= g_d && \text{on } \Gamma_B \end{aligned} \tag{B.1}$$

where $\Gamma_B \equiv \partial\Omega$. The boundary condition here, a "clamped plate" BC, is key to sustain a turbulent attractor[1].

The fourth-order operator makes the generation of solutions using a DG method non-trivial. Various authors have generated stable schemes for the Kuramoto-Sivashinsky or other fourth-order differential equations, including notably alternating upwinding on a family of auxiliary variables[2] and tailored interior penalty methods[3].

In the interest of future research using the SANS solver, we prefer to have a simple mixed form that is compatible with the lifting operator approach of Bassi and Rebay [1997] (with minimal modifications),

[1] Blonigan and Wang, 2014

[2] Yan Xu and Chi-Wang Shu. Local discontinuous Galerkin methods for the Kuramoto-Sivashinsky equations and the Ito-type coupled KdV equations. *Computer Methods in Applied Mechanics and Engineering*, 195(25):3430–3447, 2006

[3] Emmanuil H. Georgoulis, Paul Houston, and Juha Virtanen. An a posteriori error indicator for discontinuous Galerkin approximations of fourth-order elliptic problems. *IMA Journal of Numerical Analysis*, 31(1): 281–298, 09 2009

and has promise for being demonstrably dual consistent. Thus, we seek a compatible mixed form for the KSE.

Proof of dual-consistency is outside of the scope of this appendix, but our method is promising for provable dual-consistency.

We introduce a mixed variable to arrive at a coupled system of PDEs:

$$\begin{aligned} \frac{\partial u}{\partial t} + \nabla \cdot \left(cu + \frac{\alpha}{2} u^2 \right) + \nabla \cdot (\beta \nabla u + \gamma \xi \nabla a) &= f && \text{in } \Omega \\ \nabla \cdot (\nabla u) - \xi a &= 0 && \text{in } \Omega \\ u &= u_d && \text{on } \Gamma_B \\ \nabla u \cdot \hat{n} &= g_d && \text{on } \Gamma_B \end{aligned} \tag{B.2}$$

where ξ is an arbitrary constant.

This allows us to define:

$$\begin{aligned} \mathcal{F}^t &= \begin{pmatrix} u \\ 0 \end{pmatrix} & \mathcal{S} &= \begin{pmatrix} 0 \\ -\xi a \end{pmatrix} & \mathrm{f} &= \begin{pmatrix} f \\ 0 \end{pmatrix} \\ \mathcal{F}^I &= \begin{pmatrix} cu + \frac{1}{2}\alpha u^2 \\ 0 \end{pmatrix} & \mathcal{F}^V &= \begin{pmatrix} \beta \nabla u + \gamma \xi \nabla a \\ \nabla u \end{pmatrix} = -\underline{\mathbf{K}} \nabla \mathbf{u} & \underline{\mathbf{K}} &= \begin{bmatrix} -\beta & -\gamma\xi \\ -1 & 0 \end{bmatrix} \end{aligned} \tag{B.3}$$

Giving:

$$\frac{\partial \mathcal{F}^t}{\partial t} + \nabla \cdot \mathcal{F}^I(\mathbf{u}) + \nabla \cdot \mathcal{F}^V(\mathbf{u}, \nabla \mathbf{u}) + \mathcal{S}(\mathbf{u}, \nabla \mathbf{u}) = \mathrm{f} \tag{B.4}$$

where we have used $\mathbf{u} = [u, a]^\top$ to represent the complete state.

The discontinuous Galerkin form of (B.2) is given by:

$$\begin{aligned} \sum_{\kappa \in \mathcal{T}_h} \int_\kappa \mathbf{w}^\top \frac{\partial \mathcal{F}^t}{\partial t} \, \mathrm{d}\Omega + \sum_{e \in \Gamma_I} \int_e [[\mathbf{w}]]^\top \, \widehat{\mathcal{F}^I(\mathbf{u}) \cdot} \hat{n} \, \mathrm{d}\Gamma - \sum_{\kappa \in \mathcal{T}_h} \int_\kappa \nabla \mathbf{w}^\top \cdot \mathcal{F}^I(\mathbf{u}) \, \mathrm{d}\Omega \\ + \sum_{e \in \Gamma_I} \int_e [[\mathbf{w}]]^\top \, \mathcal{F}^V(\widehat{\mathbf{u}, \nabla \mathbf{u}}) \cdot \hat{n} \, \mathrm{d}\Gamma - \sum_{\kappa \in \mathcal{T}_h} \int_\kappa \nabla \mathbf{w}^\top \cdot \mathcal{F}^V(\mathbf{u}, \nabla \mathbf{u}) \, \mathrm{d}\Omega \\ + \sum_{\kappa \in \mathcal{T}_h} \int_\kappa \mathbf{w}^\top \mathcal{S}(\mathbf{u}, \nabla \mathbf{u}) \, \mathrm{d}\Omega + \sum_{e \in \Gamma_B} \int_e \mathbf{w}^\top \mathcal{F}^B(\mathbf{u}, \nabla \mathbf{u}) \cdot \hat{n} \, \mathrm{d}\Gamma = \sum_{\kappa \in \mathcal{T}_h} \int_\kappa \mathbf{w}^\top \mathrm{f} \, \mathrm{d}\Omega \end{aligned} \tag{B.5}$$

Here, $\widehat{\mathcal{F}^I(\mathbf{u}) \cdot} \hat{n}$ and $\mathcal{F}^V(\widehat{\mathbf{u}, \nabla \mathbf{u}}) \cdot \hat{n}$ are symmetric numerical fluxes and $\mathcal{F}_B(\mathbf{u}) \cdot \hat{n}$ is the numerical flux on the boundary that enforces the boundary conditions. We hope to find a choice of the boundary flux to guarantee that the system is stable. The second method of Bassi and Rebay for the viscous terms gives:

$$\begin{aligned}\sum_{\kappa\in\mathcal{T}_h}\int_\kappa \mathbf{w}^\top\frac{\partial\mathcal{F}^t}{\partial t}\,\mathrm{d}\Omega+\sum_{e\in\Gamma_I}\int_e [[\mathbf{w}]]^\top\,\widehat{\mathcal{F}^{I}(\mathbf{u})\cdot}\,\hat{n}\,\mathrm{d}\Gamma-\sum_{\kappa\in\mathcal{T}_h}\int_\kappa\nabla\mathbf{w}^\top\cdot\mathcal{F}^I(\mathbf{u})\,\mathrm{d}\Omega\\+\sum_{e\in\Gamma_I}\int_e[[\mathbf{w}]]^\top\left\{\mathcal{F}^V(\mathbf{u},\tilde{\nabla}\mathbf{u})\cdot\hat{n}\right\}\,\mathrm{d}\Gamma-\sum_{\kappa\in\mathcal{T}_h}\int_\kappa\nabla\mathbf{w}^\top\cdot\mathcal{F}^V(\mathbf{u},\tilde{\nabla}\mathbf{u})\,\mathrm{d}\Omega\\+\sum_{\kappa\in\mathcal{T}_h}\int_\kappa\mathbf{w}^\top\mathcal{S}(\mathbf{u},\tilde{\nabla}\mathbf{u})\,\mathrm{d}\Omega+\sum_{e\in\Gamma_B}\int_e\mathbf{w}^\top\mathcal{F}^B(\mathbf{u},\tilde{\nabla}\mathbf{u})\cdot\hat{n}\,\mathrm{d}\Gamma=\sum_{\kappa\in\mathcal{T}_h}\int_\kappa\mathbf{w}^\top\mathrm{f}\,\mathrm{d}\Omega,\end{aligned}\tag{B.6}$$

where the lifted gradient is given by $\tilde{\nabla}\mathbf{u}$ and will be defined in the forthcoming sections.

B.2 *Stability of steady diffusive problem*

Consider the case with $\boldsymbol{c}=\boldsymbol{\alpha}=0\implies\mathcal{F}^I\to 0$ and furthermore assume a steady system. This leaves:

$$\begin{aligned}\sum_{e\in\Gamma_I}\int_e[[\mathbf{w}]]^\top\left\{\mathcal{F}^V(\mathbf{u},\tilde{\nabla}\mathbf{u})\cdot\hat{n}\right\}\,\mathrm{d}\Gamma-\sum_{\kappa\in\mathcal{T}_h}\int_\kappa\nabla\mathbf{w}^\top\cdot\mathcal{F}^V(\mathbf{u},\tilde{\nabla}\mathbf{u})\,\mathrm{d}\Omega\\+\sum_{\kappa\in\mathcal{T}_h}\int_\kappa\mathbf{w}^\top\mathcal{S}(\mathbf{u},\tilde{\nabla}\mathbf{u})\,\mathrm{d}\Omega+\sum_{e\in\Gamma_B}\int_e\mathbf{w}^\top\mathcal{F}^V_B(\mathbf{u},\tilde{\nabla}\mathbf{u})\cdot\hat{n}\,\mathrm{d}\Gamma=\sum_{\kappa\in\mathcal{T}_h}\int_\kappa\mathbf{w}^\top\mathrm{f}\,\mathrm{d}\Omega\end{aligned}\tag{B.7}$$

Where we use a “lifted gradient” given by:

$$\tilde{\nabla}(\cdot)=\begin{cases}\nabla(\cdot)+\mathbf{R}_h\left([[\cdot]]_d\right) & \text{in } \kappa\in\mathcal{T}_h\\ \nabla(\cdot)+\eta\mathbf{r}_h^e\left([[\cdot]]_d\right) & \text{on } e\in\Gamma\end{cases}\tag{B.8}$$

The global lifting operator is defined by:

$$\begin{aligned}\sum_{\kappa\in\mathcal{T}_h}\int_\kappa\boldsymbol{\theta}\cdot\mathbf{R}_h\left([[\zeta]]_d\right)\,\mathrm{d}\Omega&=-\sum_{e\in\Gamma_I}\int_e\{\boldsymbol{\theta}\}\cdot[[\zeta]]\,\mathrm{d}\Gamma-\sum_{e\in\Gamma_B}\int_e\boldsymbol{\theta}(\zeta-\hat{\zeta})\cdot\hat{n}\,\mathrm{d}\Gamma\\\sum_{\kappa\in\mathcal{T}_h}\int_\kappa\boldsymbol{\theta}\cdot\mathbf{R}_h^0([[\zeta]])\,\mathrm{d}\Omega&=-\sum_{e\in\Gamma_I}\int_e\{\boldsymbol{\theta}\}\cdot[[\zeta]]\,\mathrm{d}\Gamma-\sum_{e\in\Gamma_B}\int_e\boldsymbol{\theta}\zeta\cdot\hat{n}\,\mathrm{d}\Gamma\end{aligned}\tag{B.9}$$

where $\hat{\zeta}=u_d$ when $\zeta=u$ and $\hat{\zeta}=a$ when $\zeta=a$; this implies:

$$\sum_{\kappa\in\mathcal{T}_h}\int_\kappa\boldsymbol{\theta}\cdot\mathbf{R}_h\left([[\zeta]]_d\right)\,\mathrm{d}\Omega=\sum_{\kappa\in\mathcal{T}_h}\int_\kappa\boldsymbol{\theta}\cdot\mathbf{R}_h^0([[\zeta]])\,\mathrm{d}\Omega+\sum_{e\in\Gamma_B}\int_e\boldsymbol{\theta}\hat{\zeta}\cdot\hat{n}\,\mathrm{d}\Gamma\tag{B.10}$$

where $\hat{\zeta}$ can be replaced with a Dirichlet boundary state if it exists.
The local lifting operator is, similarly, defined by:

$$
\begin{aligned}
\sum_{\kappa \in \{\kappa\}^e} \int_\kappa \boldsymbol{\theta} \cdot \mathbf{r}_h^e\left([[u]]_d\right) \, \mathrm{d}\Omega &= \begin{cases} -\int_e \{\boldsymbol{\theta}\} \cdot [[u]] \, \mathrm{d}\Gamma & \text{if } e \in \Gamma_I \\ -\int_e \{\boldsymbol{\theta}\} \cdot (u - u_d) \, \mathrm{d}\Gamma & \text{if } e \in \Gamma_B \\ 0 & \text{otherwise} \end{cases} \\
\sum_{\kappa \in \{\kappa\}^e} \int_\kappa \boldsymbol{\theta} \cdot \mathbf{r}_h^e\left([[a]]_d\right) \, \mathrm{d}\Omega &= \begin{cases} -\int_e \{\boldsymbol{\theta}\} \cdot [[u]] \, \mathrm{d}\Gamma & \text{if } e \in \Gamma_I \\ 0 & \text{otherwise} \end{cases},
\end{aligned}
\tag{B.11}
$$

and

$$
\sum_{\kappa \in \{\kappa\}^e} \int_\kappa \boldsymbol{\theta} \cdot \mathbf{r}_{h,0}^e([[\zeta]]) \, \mathrm{d}\Omega = \begin{cases} -\int_e \{\boldsymbol{\theta}\} \cdot [[\zeta]] \, \mathrm{d}\Gamma & \text{if } e \in \Gamma_I \\ -\int_e \boldsymbol{\theta}\zeta \cdot \hat{n} \, \mathrm{d}\Gamma & \text{if } e \in \Gamma_B \end{cases}
\tag{B.12}
$$

such that:

$$
\begin{aligned}
\sum_{\kappa \in \{\kappa\}^e} \int_\kappa \boldsymbol{\theta} \cdot \mathbf{r}_h^e\left([[u]]_d\right) \, \mathrm{d}\Omega &= \sum_{\kappa \in \{\kappa\}^e} \int_\kappa \boldsymbol{\theta} \cdot \mathbf{r}_{h,0}^e([[u]]) \, \mathrm{d}\Omega + \int_e \boldsymbol{\theta} u_d \cdot \hat{n} \, \mathrm{d}\Gamma \\
\sum_{\kappa \in \{\kappa\}^e} \int_\kappa \boldsymbol{\theta} \cdot \mathbf{r}_h^e\left([[a]]_d\right) \, \mathrm{d}\Omega &= \sum_{\kappa \in \{\kappa\}^e} \int_\kappa \boldsymbol{\theta} \cdot \mathbf{r}_{h,0}^e([[a]]) \, \mathrm{d}\Omega + \int_e \boldsymbol{\theta} a \cdot \hat{n} \, \mathrm{d}\Gamma
\end{aligned}
\tag{B.13}
$$

Now, we can substitute the definitions into the equation we want to use to prove coercivity:

$$
\begin{aligned}
&\sum_{e \in \Gamma_I} \int_e [[w]] \left\{ \left[\beta \tilde{\nabla} u + \gamma \tilde{\nabla} a\right] \cdot \hat{n} \right\} \, \mathrm{d}\Gamma + \sum_{e \in \Gamma_I} \int_e [[b]] \left\{ \tilde{\nabla} u \cdot \hat{n} \right\} \, \mathrm{d}\Gamma \\
&\quad - \sum_{\kappa \in \mathcal{T}_h} \int_\kappa \nabla w \cdot \left[\beta \tilde{\nabla} u + \gamma \tilde{\nabla} a\right] \, \mathrm{d}\Omega - \sum_{\kappa \in \mathcal{T}_h} \int_\kappa \nabla b \cdot \tilde{\nabla} u \, \mathrm{d}\Omega \\
&\qquad - \sum_{\kappa \in \mathcal{T}_h} \int_\kappa ba \, \mathrm{d}\Omega + \sum_{e \in \Gamma_B} \int_e \begin{pmatrix} w \\ b \end{pmatrix}^\top \mathcal{F}_B^V\left(\mathbf{u}, \tilde{\nabla}\mathbf{u}\right) \cdot \hat{n} \, \mathrm{d}\Gamma = \sum_{\kappa \in \mathcal{T}_h} \int_\kappa wf \, \mathrm{d}\Omega
\end{aligned}
\tag{B.14}
$$

On the boundary, the boundary flux is given by:

$$
\mathcal{F}_B^V \cdot \hat{n} = \begin{pmatrix} \beta \tilde{\nabla} u \cdot \hat{n} + \gamma \tilde{\nabla} a \cdot \hat{n} \\ g_d \end{pmatrix}
\tag{B.15}
$$

Here, we have replaced the Neumann boundary condition on the second equation. This gives:

$$
\begin{aligned}
&\sum_{e\in\Gamma_I}\int_e [[w]]\left\{\left[\beta\left(\nabla u+\eta\mathbf{r}_h^e\left([[u]]_d\right)\right)+\gamma\left(\nabla a+\eta\mathbf{r}_h^e\left([[a]]_d\right)\right)\right]\cdot\hat{n}\right\}\ \mathrm{d}\Gamma\\
&\qquad+\sum_{e\in\Gamma_I}\int_e [[b]]\left\{\left(\nabla u+\eta\mathbf{r}_h^e\left([[u]]_d\right)\right)\cdot\hat{n}\right\}\ \mathrm{d}\Gamma\\
&-\sum_{\kappa\in\mathcal{T}_h}\int_\kappa \nabla w\cdot\left[\beta\left(\nabla u+\mathbf{R}_h\left([[u]]_d\right)\right)+\gamma\left(\nabla a+\mathbf{R}_h\left([[a]]_d\right)\right)\right]\ \mathrm{d}\Omega\\
&\qquad-\sum_{\kappa\in\mathcal{T}_h}\int_\kappa \nabla b\cdot\left(\nabla u+\mathbf{R}_h\left([[u]]_d\right)\right)\ \mathrm{d}\Omega-\sum_{\kappa\in\mathcal{T}_h}\int_\kappa ba\ \mathrm{d}\Omega\\
&\quad+\sum_{e\in\Gamma_B}\int_e w\left[\beta\left(\nabla u+\eta\mathbf{r}_h^e\left([[u]]_d\right)\right)+\gamma\left(\nabla a+\eta\mathbf{r}_h^e\left([[a]]_d\right)\right)\right]\cdot\hat{n}\ \mathrm{d}\Gamma=-\sum_{e\in\Gamma_B}\int_e bg_d\ \mathrm{d}\Gamma\\
&\qquad\qquad+\sum_{\kappa\in\mathcal{T}_h}\int_\kappa wf\ \mathrm{d}\Omega
\end{aligned}
\tag{B.16}
$$

Consider testing with $\mathbf{w}=[u,-\gamma a]$:

$$
\begin{aligned}
&\sum_{\kappa\in\mathcal{T}_h}\int_\kappa \gamma a^2\ \mathrm{d}\Omega+\sum_{\kappa\in\mathcal{T}_h}\int_\kappa \gamma\nabla a\cdot\nabla u\ \mathrm{d}\Omega-\sum_{\kappa\in\mathcal{T}_h}\int_\kappa \gamma\nabla u\cdot\nabla a\ \mathrm{d}\Omega\\
&-\sum_{\kappa\in\mathcal{T}_h}\int_\kappa \gamma\nabla u\cdot\mathbf{R}_h\left([[a]]_d\right)\ \mathrm{d}\Omega+\sum_{\kappa\in\mathcal{T}_h}\int_\kappa \gamma\nabla a\cdot\mathbf{R}_h\left([[u]]_d\right)\ \mathrm{d}\Omega\\
&+\sum_{e\in\Gamma_I}\int_e \gamma\,[[u]]\left\{\nabla a\cdot\hat{n}\right\}\ \mathrm{d}\Gamma+\sum_{e\in\Gamma_I}\int_e \gamma\eta\,[[u]]\left\{\mathbf{r}_h^e\left([[a]]_d\right)\cdot\hat{n}\right\}\ \mathrm{d}\Gamma\\
&-\sum_{e\in\Gamma_I}\int_e \gamma\,[[a]]\left\{\nabla u\cdot\hat{n}\right\}\ \mathrm{d}\Gamma-\sum_{e\in\Gamma_I}\int_e \gamma\eta\,[[a]]\left\{\mathbf{r}_h^e\left([[u]]_d\right)\cdot\hat{n}\right\}\ \mathrm{d}\Gamma\\
&\qquad+\sum_{e\in\Gamma_B}\int_e \gamma u\nabla a\cdot\hat{n}\ \mathrm{d}\Gamma+\sum_{e\in\Gamma_B}\int_e \gamma\eta u\mathbf{r}_h^e\left([[a]]_d\right)\cdot\hat{n}\ \mathrm{d}\Gamma\\
&\qquad-\sum_{\kappa\in\mathcal{T}_h}\int_\kappa \beta\nabla u\cdot\nabla u\ \mathrm{d}\Omega-\sum_{\kappa\in\mathcal{T}_h}\int_\kappa \beta\nabla u\cdot\mathbf{R}_h\left([[u]]_d\right)\ \mathrm{d}\Omega\\
&+\sum_{e\in\Gamma_I}\int_e \beta\,[[u]]\left\{\nabla u\cdot\hat{n}\right\}\ \mathrm{d}\Gamma+\sum_{e\in\Gamma_I}\int_e \beta\eta\,[[u]]\left\{\mathbf{r}_h^e\left([[u]]_d\right)\cdot\hat{n}\right\}\ \mathrm{d}\Gamma\\
&\qquad\qquad+\sum_{e\in\Gamma_B}\int_e \beta u\nabla u\cdot\hat{n}\ \mathrm{d}\Gamma+\sum_{e\in\Gamma_B}\int_e \beta\eta u\mathbf{r}_h^e\left([[u]]_d\right)\cdot\hat{n}\ \mathrm{d}\Gamma=\sum_{e\in\Gamma_B}\int_e bg_d\ \mathrm{d}\Gamma\\
&\qquad\qquad\qquad+\sum_{\kappa\in\mathcal{T}_h}\int_\kappa wf\ \mathrm{d}\Omega
\end{aligned}
\tag{B.17}
$$

Noting that $\beta<0$ for stable second-order system, we find that the blue term is positive for $\gamma>0$, the cyan term is positive for $\gamma>0$ and $\beta<0$, magenta terms can be shown to be positive following the DGBR2 Dirichlet problem proof[4], and red terms cancel out or go to zero (note here that $\mathbf{r}_h^e\left([[a]]_d\right)\to 0$ when $e\in\Gamma_B$).

[4]

In order to prove coercivity, we must show that the remaining terms, in black, are positive; in sum:

$$\begin{aligned}
\Phi \equiv &- \sum_{\kappa\in\mathcal{T}_h} \int_\kappa \gamma \nabla u \cdot \mathbf{R}_h \left([[a]]_d\right) \, \mathrm{d}\Omega \, + \sum_{\kappa\in\mathcal{T}_h} \int_\kappa \gamma \nabla a \cdot \mathbf{R}_h \left([[u]]_d\right) \, \mathrm{d}\Omega \\
&+ \sum_{e\in\Gamma_I} \int_e \gamma \, [[u]] \cdot \{\nabla a\} \, \mathrm{d}\Gamma + \sum_{e\in\Gamma_I} \int_e \gamma\eta \, [[u]] \cdot \{\mathbf{r}_h^e \left([[a]]_d\right)\} \, \mathrm{d}\Gamma \\
&- \sum_{e\in\Gamma_I} \int_e \gamma \, [[a]] \cdot \{\nabla u\} \, \mathrm{d}\Gamma - \sum_{e\in\Gamma_I} \int_e \gamma\eta \, [[a]] \cdot \{\mathbf{r}_h^e \left([[u]]_d\right)\} \, \mathrm{d}\Gamma \\
&+ \sum_{e\in\Gamma_B} \int_e \gamma u \nabla a \cdot \hat{n} \, \mathrm{d}\Gamma
\end{aligned} \tag{B.18}$$

Using $\boldsymbol{\theta} = \nabla u$ and $\zeta = a$ and $\boldsymbol{\theta} = \nabla a$ and $\zeta = u$, respectively:

$$\begin{aligned}
\sum_{\kappa\in\mathcal{T}_h} \int_\kappa \nabla u \cdot \mathbf{R}_h \left([[a]]_d\right) \, \mathrm{d}\Omega &= - \sum_{e\in\Gamma_I} \int_e \{\nabla u\} \cdot [[a]] \, \mathrm{d}\Gamma \\
\sum_{\kappa\in\mathcal{T}_h} \int_\kappa \nabla a \cdot \mathbf{R}_h \left([[u]]_d\right) \, \mathrm{d}\Omega &= - \sum_{e\in\Gamma_I} \int_e \{\nabla a\} \cdot [[u]] \, \mathrm{d}\Gamma - \sum_{e\in\Gamma_B} \int_e \nabla a u \cdot \hat{n} \, \mathrm{d}\Gamma + \sum_{e\in\Gamma_B} \int_e \nabla a u_d \cdot \hat{n} \, \mathrm{d}\Gamma
\end{aligned} \tag{B.19}$$

Substituting:

$$\begin{aligned}
\Phi = &\sum_{e\in\Gamma_I} \int_e \gamma \{\nabla u\} \cdot [[a]] \, \mathrm{d}\Gamma - \sum_{e\in\Gamma_I} \int_e \gamma \{\nabla a\} \cdot [[u]] \, \mathrm{d}\Gamma - \sum_{e\in\Gamma_B} \int_e \gamma \nabla a u \cdot \hat{n} \, \mathrm{d}\Gamma + \sum_{e\in\Gamma_B} \int_e \gamma \nabla a u_d \cdot \hat{n} \, \mathrm{d}\Gamma \\
&+ \sum_{e\in\Gamma_I} \int_e \gamma \, [[u]] \cdot \{\nabla a\} \, \mathrm{d}\Gamma + \sum_{e\in\Gamma_I} \int_e \gamma\eta \, [[u]] \cdot \{\mathbf{r}_h^e \left([[a]]_d\right)\} \, \mathrm{d}\Gamma \\
&- \sum_{e\in\Gamma_I} \int_e \gamma \, [[a]] \cdot \{\nabla u\} \, \mathrm{d}\Gamma - \sum_{e\in\Gamma_I} \int_e \gamma\eta \, [[a]] \cdot \{\mathbf{r}_h^e \left([[u]]_d\right)\} \, \mathrm{d}\Gamma \\
&+ \sum_{e\in\Gamma_B} \int_e \gamma w \nabla a \cdot \hat{n} \, \mathrm{d}\Gamma
\end{aligned} \tag{B.20}$$

As previously, red terms cancel out. The cyan term here, noting that this term emerges from the $b = \gamma a$ transformation, is reducible to a constant RHS term and we can ignore it while proving coercivity.

Now, for the remaining terms, we use $\boldsymbol{\theta} = \mathbf{r}_h^e \left([[a]]_d\right)$ and $\boldsymbol{\theta} = \mathbf{r}_h^e \left([[u]]_d\right)$, respectively, to find:

$$\sum_{\kappa\in\{\kappa\}^e}\int_\kappa \mathbf{r}_h^e\left([[a]]_d\right)\cdot\mathbf{r}_h^e\left([[u]]_d\right)\,\mathrm{d}\Omega=\begin{cases}-\int_e\left\{\mathbf{r}_h^e\left([[a]]_d\right)\right\}\cdot[[u]]\,\mathrm{d}\Gamma & \text{if } e\in\Gamma_I\\ -\int_e\left\{\mathbf{r}_h^e\left([[a]]_d\right)\right\}\cdot\left(u-u_d\right)\mathrm{d}\Gamma & \text{if } e\in\Gamma_B\\ 0 & \text{otherwise}\end{cases}$$
$$\sum_{\kappa\in\{\kappa\}^e}\int_\kappa \mathbf{r}_h^e\left([[u]]_d\right)\cdot\mathbf{r}_h^e\left([[a]]_d\right)\,\mathrm{d}\Omega=\begin{cases}-\int_e\left\{\mathbf{r}_h^e\left([[u]]_d\right)\right\}\cdot[[u]]\,\mathrm{d}\Gamma & \text{if } e\in\Gamma_I\\ 0 & \text{otherwise}\end{cases}\tag{B.21}$$

By summing over the interior edges, we find:

$$\Phi=-\gamma\eta\sum_{e\in\Gamma_I}\sum_{\kappa\in\{\kappa\}^e}\int_\kappa\mathbf{r}_h^e\left([[a]]_d\right)\cdot\mathbf{r}_h^e\left([[u]]_d\right)\,\mathrm{d}\Omega+\gamma\eta\sum_{e\in\Gamma_I}\sum_{\kappa\in\{\kappa\}^e}\int_\kappa\mathbf{r}_h^e\left([[u]]_d\right)\cdot\mathbf{r}_h^e\left([[a]]_d\right)\,\mathrm{d}\Omega\tag{B.22}$$

Thus, $\Phi=0$, and L_2 stability follows for $\gamma>0$ and $\beta<0$.

In the case that $\beta\to0$, we find that the formulation does not have a unique solution because there are infinitely many solutions, varying u, at a given minimizer of a; this is equivalent to the lack of an inf-sup condition on the Ciarlet-Raviart saddle point problem. We now seek to extend the BR2 stabilization to guarantee the existence of a solution.

B.3 *Extended DGBR2 formulation for biharmonic operators*

In (B.8), the formulation of the lifted gradient comes from the idea of adding a consistent stabilization to penalize jumps in the solution. Now, we consider such a penalty term that is now independent of β but instead dependent on γ:

$$\begin{aligned}\mathcal{R}^s(w,u)&=(s\gamma)\sum_{e\in\Gamma_I^0}\sum_{\kappa\in\mathcal{T}_h}\int_\kappa\mathbf{r}_{h,0}^e([[w]])\cdot\mathbf{r}_h^e\left([[u]]_d\right)\,\mathrm{d}\Omega\\&=(s\gamma)\sum_{e\in\Gamma_I^0}\sum_{\kappa\in\mathcal{T}_h}\int_\kappa\mathbf{r}_{h,0}^e([[w]])\cdot\mathbf{r}_{h,0}^e([[u]])\,\mathrm{d}\Omega+(s\gamma)\sum_{e\in\Gamma_B}\int_e\mathbf{r}_{h,0}^e([[w]])u_d\cdot\hat{n}\,\mathrm{d}\Gamma\\&=-(s\gamma)\sum_{e\in\Gamma_I}\int_e[[w]]\cdot\mathbf{r}_h^e\left([[u]]_d\right)\,\mathrm{d}\Gamma-(s\gamma)\sum_{e\in\Gamma_B}\int_e w\mathbf{r}_h^e\left([[u]]_d\right)\cdot\hat{n}\,\mathrm{d}\Gamma\end{aligned}\tag{B.23}$$

These relations follow straightforwardly from the definitions. We can see that adding this stabilization term is equivalent to adding an additional penalty to the boundary jumps. Two facts are trivially observable: first, that by taking $w=u$ this term is coercive in u; and second, because each term is dependent on the jump in the solution variable, it is zero for sufficiently smooth solutions.

We can add this term to (B.16):

$$
\begin{aligned}
&\sum_{e\in\Gamma_I}\int_e [[w]]\left\{\beta\left(\nabla u+\eta\mathbf{r}_h^e\left([[u]]_d\right)\right)\cdot\hat{n}\right\}\,\mathrm{d}\Gamma \\
&+\sum_{e\in\Gamma_I}\int_e [[w]]\left\{\gamma\left(\nabla a+\eta\mathbf{r}_h^e\left([[a]]_d\right)-s\mathbf{r}_h^e\left([[u]]_d\right)\right)\cdot\hat{n}\right\}\,\mathrm{d}\Gamma \\
&-\sum_{e\in\Gamma_I}\int_e [[b]]\left\{\left(\nabla u+\eta\mathbf{r}_h^e\left([[u]]_d\right)\right)\cdot\hat{n}\right\}\,\mathrm{d}\Gamma \\
&-\sum_{\kappa\in\mathcal{T}_h}\int_\kappa \nabla w\cdot\left[\beta\left(\nabla u+\mathbf{R}_h\left([[u]]_d\right)\right)+\gamma\left(\nabla a+\mathbf{R}_h\left([[a]]_d\right)\right)\right]\,\mathrm{d}\Omega \\
&+\sum_{\kappa\in\mathcal{T}_h}\int_\kappa \nabla b\cdot\left(\nabla u+\mathbf{R}_h\left([[u]]_d\right)\right)\,\mathrm{d}\Omega+\sum_{\kappa\in\mathcal{T}_h}\int_\kappa ba\,\mathrm{d}\Omega \\
&+\sum_{e\in\Gamma_B}\int_e w\left(\beta\left(\nabla u+\eta\mathbf{r}_h^e\left([[u]]_d\right)-s\mathbf{r}_h^e\left([[u]]_d\right)\right)\cdot\hat{n}\right)\,\mathrm{d}\Gamma \\
&+\sum_{e\in\Gamma_B}\int_e w\left(\gamma\left(\nabla a+\eta\mathbf{r}_h^e\left([[a]]_d\right)\right)\cdot\hat{n}\right)\,\mathrm{d}\Gamma=\sum_{e\in\Gamma_B}\int_e bg_d\,\mathrm{d}\Gamma \\
&\qquad+\sum_{\kappa\in\mathcal{T}_h}\int_\kappa wf\,\mathrm{d}\Omega
\end{aligned}
\tag{B.24}
$$

This implies a new approach to the lifted gradients from before; instead of stabilizing as we have previously, we now introduce a stabilization matrix to the lifted gradients:

$$
\underline{\mathbf{A}}_{\text{stab}}=\begin{bmatrix}\eta & 0\\ -s & \eta\end{bmatrix}
\tag{B.25}
$$

and now:

$$
\begin{pmatrix}\tilde{\nabla}u\\ \tilde{\nabla}a\end{pmatrix}=\begin{cases}\begin{pmatrix}\nabla u\\ \nabla a\end{pmatrix}+\begin{pmatrix}[[u]]_d\\ [[a]]_d\end{pmatrix} & \text{in } \kappa\in\mathcal{T}_h\\ \begin{pmatrix}\nabla u\\ \nabla a\end{pmatrix}+\underline{\mathbf{A}}_{\text{stab}}\begin{pmatrix}[[u]]_d\\ [[a]]_d\end{pmatrix} & \text{on } e\in\Gamma\end{cases}
\tag{B.26}
$$

which should now guarantee stability and coercivity of the Ciarlet-Raviart DGBR2 mixed form for biharmonic operators for all stable choices of $\gamma > 0$. For a shorthand, we refer to this stability form as the "DGBR4" method.

B.4 DIRK-DAE time marching for DGBR4

Now, consider a generalized semi-discrete form for unsteady DG discretizations:

$$\begin{aligned}\sum_{\kappa\in\mathcal{T}_h}\int_{\kappa} w\frac{\partial}{\partial t}\left(\mathcal{F}_u^t(u(q),a)\right)\,\mathrm{d}\Omega+\mathcal{R}_u^{\mathrm{semi}}(w;u(q),a)&=0\\ \sum_{\kappa\in\mathcal{T}_h}\int_{\kappa} b\frac{\partial}{\partial t}\left(\mathcal{F}_a^t(u(q),a)\right)\,\mathrm{d}\Omega+\mathcal{R}_a^{\mathrm{semi}}(b;u(q),a)&=0.\end{aligned} \tag{B.27}$$

This can be specialized to the KSE problem by taking the result of (B.6) with the DGBR4 definition of the lifting operator,

When we apply timestepping methods to problems of this type, w and b will be exclusively dependent on space, not time, so $w = w(\vec{x})$ and $b = b(\vec{x})$. Thus we will want to approximate

$$\begin{aligned}\frac{\partial}{\partial t}\left(\sum_{\kappa\in\mathcal{T}_h}\int_{\kappa} w\mathcal{F}_u^t(u(q),a)\,\mathrm{d}\Omega\right)+\mathcal{R}_u^{\mathrm{semi}}(w;u(q),a)&=0\\ \frac{\partial}{\partial t}\left(\sum_{\kappa\in\mathcal{T}_h}\int_{\kappa} b\mathcal{F}_a^t(u(q),a)\,\mathrm{d}\Omega\right)+\mathcal{R}_a^{\mathrm{semi}}(b;u(q),a)&=0\end{aligned} \tag{B.28}$$

$\forall w, b \in \mathcal{V}$ with an appropriate timestepped form.

Given this form, we also note that for the Galerkin methods, we represent the solutions q and a by a summation of degrees of freedom $Q_k = [\mathbf{Q}]_k$ and $A_k = [\mathbf{A}]_k$ over a set of N_{basis} basis functions $\phi_k(\vec{x})$:

$$\begin{aligned}q = q(\vec{x}) &= \sum_{k=1}^{N_{\text{basis}}} Q_k\phi_k(\vec{x}) = q(\mathbf{Q})\\ a = a(\vec{x}) &= \sum_{k=1}^{N_{\text{basis}}} A_k\phi_k(\vec{x}) = a(\mathbf{A})\end{aligned} \tag{B.29}$$

Now given a test function w_j, we can write an equivalent form of (B.28) for the semidiscrete system variables $\mathbf{Q} = \mathbf{Q}(t)$ and $\mathbf{A} = \mathbf{A}(t)$:

$$\begin{aligned}\frac{\partial}{\partial t}\left(\sum_{\kappa\in\mathcal{T}_h}\int_{\kappa} w_j\mathcal{F}_u^t(\mathbf{Q},\mathbf{A})\,\mathrm{d}\Omega\right)+\mathcal{R}_u^{\mathrm{semi}}(w_j;\mathbf{Q},\mathbf{A})&=0\\ \frac{\partial}{\partial t}\left(\sum_{\kappa\in\mathcal{T}_h}\int_{\kappa} b_j\mathcal{F}_a^t(\mathbf{Q},\mathbf{A})\,\mathrm{d}\Omega\right)+\mathcal{R}_a^{\mathrm{semi}}(b_j;\mathbf{Q},\mathbf{A})&=0\end{aligned} \tag{B.30}$$

where the system fluxes here are defined by forwarding the DOF vectors to the relevant functions.

For low dispersive errors, we would like to use a diagonally implicit RK approach. In this case, we solve:

$$\begin{aligned}\frac{1}{\Delta t}\mathcal{R}_u^t(w_j;\mathbf{Q}^{n+1},\mathbf{A}^{n+1}) &= \frac{1}{\Delta t}\mathcal{R}_u^t(w_j;\mathbf{Q}^{n},\mathbf{A}^{n}) + \sum_{k=1}^{s}\beta_k\mathbf{K}_j^{(k)}(\mathbf{Q}^{(k)},\mathbf{A}^{(k)})\\ \frac{1}{\Delta t}\mathcal{R}_a^t(b_j;\mathbf{Q}^{n+1},\mathbf{A}^{n+1}) &= \frac{1}{\Delta t}\mathcal{R}_a^t(b_j;\mathbf{Q}^{n},\mathbf{A}^{n}) + \sum_{k=1}^{s}\beta_k\mathbf{L}_j^{(k)}(\mathbf{Q}^{(k)},\mathbf{A}^{(k)})\end{aligned} \tag{B.31}$$

Where each of the stage solutions $\mathbf{Q}^{(k)}$ and $\mathbf{A}^{(k)}$ are solutions to:

$$\begin{aligned}\mathbf{K}_j^{(k)}(\mathbf{Q}^{(k)},\mathbf{A}^{(k)}) &\equiv \frac{1}{\Delta t}\mathcal{R}_u^t(w_j;\mathbf{Q}^{(k)},\mathbf{A}^{(k)}) - \frac{1}{\Delta t}\mathcal{R}_u^t(w_j;\mathbf{Q}^{n},\mathbf{A}^{n})\\ &= -\sum_{m=1}^{k}\alpha_{km}\mathcal{R}_u^{\text{semi}}(w_j;\mathbf{Q}^{(m)},\mathbf{A}^{(m)})\\ \mathbf{L}_j^{(k)}(\mathbf{Q}^{(k)},\mathbf{A}^{(k)}) &\equiv \frac{1}{\Delta t}\mathcal{R}_a^t(b_j;\mathbf{Q}^{n},\mathbf{A}^{n}) - \frac{1}{\Delta t}\mathcal{R}_a^t(b_j;\mathbf{Q}^{(k)},\mathbf{A}^{(k)})\\ &= -\sum_{m=1}^{k}\alpha_{km}\mathcal{R}_a^{\text{semi}}(b_j;\mathbf{Q}^{(m)},\mathbf{A}^{(m)}).\end{aligned} \tag{B.32}$$

for all j such that $\mathrm{K}_j^{(k)} = [\mathbf{K}^{(k)}]_j$ and likewise $\mathrm{L}_j^{(k)} = [\mathbf{L}^{(k)}]_j$. Thus, we can solve a problem of the form:

$$\begin{aligned}\mathrm{B}^u_{\mathrm{RK}(k),j}(\mathbf{Q}^{(k)},\mathbf{A}^{(k)}) &\equiv \frac{1}{\alpha_{kk}\Delta t}\mathcal{R}_u^t(w_j;\mathbf{Q}^{(k)},\mathbf{A}^{(k)}) + \mathcal{R}_u^{\text{semi}}(w_j;\mathbf{Q}^{(k)},\mathbf{A}^{(k)})\\ &= \frac{1}{\alpha_{kk}\Delta t}\mathcal{R}_u^t(w_j;\mathbf{Q}^{n},\mathbf{A}^{n}) - \sum_{m=1}^{k-1}\frac{\alpha_{km}}{\alpha_{kk}}\mathcal{R}_u^{\text{semi}}(w_j;\mathbf{Q}^{(m)},\mathbf{A}^{(m)}) \equiv \mathrm{f}^u_{\mathrm{RK}(k),j}\\ \mathrm{B}^a_{\mathrm{RK}(k),j}(\mathbf{Q}^{(k)},\mathbf{A}^{(k)}) &\equiv \frac{1}{\alpha_{kk}\Delta t}\mathcal{R}_a^t(b_j;\mathbf{Q}^{(k)},\mathbf{A}^{(k)}) + \mathcal{R}_a^{\text{semi}}(w_j;\mathbf{Q}^{(k)},\mathbf{A}^{(k)})\\ &= \frac{1}{\alpha_{kk}\Delta t}\mathcal{R}_a^t(b_j;\mathbf{Q}^{n},\mathbf{A}^{n}) - \sum_{m=1}^{k-1}\frac{\alpha_{km}}{\alpha_{kk}}\mathcal{R}_a^{\text{semi}}(b_j;\mathbf{Q}^{(m)},\mathbf{A}^{(m)}) \equiv \mathrm{f}^a_{\mathrm{RK}(k),j}.\end{aligned} \tag{B.33}$$

If solve this system using a Newton-like method, we will do it by generating a Jacobian like:

$$\frac{\partial(\mathrm{B}^u_{\mathrm{RK}(k),j},\mathrm{B}^a_{\mathrm{RK}(k),j})}{\partial(\mathbf{Q}^{(k)},\mathbf{A}^{(k)})} = \begin{bmatrix}\left(\frac{1}{\alpha_{kk}\Delta t}\frac{\partial\mathcal{R}_u^t}{\partial\mathbf{Q}} + \frac{\partial\mathcal{R}_u^{\text{semi}}}{\partial\mathbf{Q}}\right)\Big|_{w=w_j} & \left(\frac{1}{\alpha_{kk}\Delta t}\frac{\partial\mathcal{R}_u^t}{\partial\mathbf{A}} + \frac{\partial\mathcal{R}_u^{\text{semi}}}{\partial\mathbf{A}}\right)\Big|_{w=w_j}\\ \left(\frac{1}{\alpha_{kk}\Delta t}\frac{\partial\mathcal{R}_a^t}{\partial\mathbf{Q}} + \frac{\partial\mathcal{R}_a^{\text{semi}}}{\partial\mathbf{Q}}\right)\Big|_{w=w_j} & \left(\frac{1}{\alpha_{kk}\Delta t}\frac{\partial\mathcal{R}_a^t}{\partial\mathbf{A}} + \frac{\partial\mathcal{R}_a^{\text{semi}}}{\partial\mathbf{A}}\right)\Big|_{w=w_j}\end{bmatrix} \tag{B.34}$$

This Jacobian should remain nonsingular for $\mathcal{F}_a^t \rightarrow 0$ so long as the spatial problem $\mathcal{R}_a^{\text{semi}}$ remains nonsingular. Thus, we can solve for all of the intermediate stage working states $\mathbf{Q}^{(k)}$ and $\mathbf{A}^{(k)}$. In turn, we can evaluate the stage-wise conservative steps $\mathbf{K}^{(k)}(\mathbf{Q}^{(k)}, \mathbf{A}^{(k)})$ and $\mathbf{L}^{(k)}(\mathbf{Q}^{(k)}, \mathbf{A}^{(k)})$.

With these solutions in hand, now seek to use (B.31) to update the solution for $\mathbf{Q}^{n+1}$ and $\mathbf{A}^{n+1}$ at t^{n+1}. In order to do so, we now seek to solve a system of the form:

$$\begin{aligned} \mathrm{B}^{u}_{\mathrm{RK},j}(\mathbf{Q}^{n+1}, \mathbf{A}^{n+1}) &\equiv \frac{1}{\Delta t}\mathcal{R}^{t}_{u}(w_j; \mathbf{Q}^{n+1}, \mathbf{A}^{n+1}) \\ &= \frac{1}{\Delta t}\mathcal{R}^{t}_{u}(w_j; \mathbf{Q}^{n}, \mathbf{A}^{n}) - \sum_{k=1}^{s} \beta_k \mathrm{K}^{(k)}_{j}(\mathbf{Q}^{(k)}, \mathbf{A}^{(k)}) \equiv \mathrm{f}^{u}_{\mathrm{RK},j} \\ \mathrm{B}^{a}_{\mathrm{RK},j}(\mathbf{Q}^{n+1}, \mathbf{A}^{n+1}) &\equiv \frac{1}{\Delta t}\mathcal{R}^{t}_{a}(b_j; \mathbf{Q}^{n}, \mathbf{A}^{n}) \\ &= \frac{1}{\Delta t}\mathcal{R}^{t}_{a}(b_j; \mathbf{Q}^{n}, \mathbf{A}^{n}) - \sum_{k=1}^{s} \beta_k \mathrm{L}^{(k)}_{j}(\mathbf{Q}^{(k)}, \mathbf{A}^{(k)}) \equiv \mathrm{f}^{a}_{\mathrm{RK},j}. \end{aligned} \tag{B.35}$$

Using a Newton-like method, we find a Jacobian like:

$$\frac{\partial(\mathrm{B}^{u}_{\mathrm{RK},j}, \mathrm{B}^{a}_{\mathrm{RK},j})}{\partial(\mathbf{Q}^{n+1}, \mathbf{A}^{n+1})} = \begin{bmatrix} \left(\frac{1}{\Delta t}\frac{\partial \mathcal{R}^{t}_{u}}{\partial \mathbf{Q}}\right)\Big|_{w=w_j} & \left(\frac{1}{\Delta t}\frac{\partial \mathcal{R}^{t}_{u}}{\partial \mathbf{A}}\right)\Big|_{w=w_j} \\ \left(\frac{1}{\Delta t}\frac{\partial \mathcal{R}^{t}_{a}}{\partial \mathbf{Q}}\right)\Big|_{w=w_j} & \left(\frac{1}{\Delta t}\frac{\partial \mathcal{R}^{t}_{a}}{\partial \mathbf{A}}\right)\Big|_{w=w_j} \end{bmatrix}, \tag{B.36}$$

which will be singular when $\mathcal{R}^{t}_{a}$ (or $\mathcal{R}^{t}_{u}$ for that matter) goes to zero. In that case, (B.30) is an index-1 set of differential-algebraic equations (DAEs), rather than a system of ordinary differential equations (ODEs). To handle this, we specialize the methods of **?** for our working/conservative variable transformations. Thus, we want to advance $\mathbf{Q}$ using the Runge-Kutta method on $\mathcal{R}^{t}_{u}$ while advancing $\mathbf{A}$ by solving $\mathcal{R}^{t}_{a}(\mathbf{Q}^{n+1}, \mathbf{A}^{n+1}) = 0$. This gives an augmented system for the case $\mathcal{R}^{t}_{a} \to 0$:

$$\begin{aligned} \mathrm{B}^{u}_{\mathrm{RK},j}(\mathbf{Q}^{n+1}, \mathbf{A}^{n+1}) &\equiv \frac{1}{\Delta t}\mathcal{R}^{t}_{u}(w_j; \mathbf{Q}^{n+1}, \mathbf{A}^{n+1}) \\ &= \frac{1}{\Delta t}\mathcal{R}^{t}_{u}(w_j; \mathbf{Q}^{n}, \mathbf{A}^{n}) - \sum_{k=1}^{s} \beta_k \mathrm{K}^{(k)}_{j}(\mathbf{Q}^{(k)}, \mathbf{A}^{(k)}) \equiv \mathrm{f}^{u}_{\mathrm{RK},j} \\ \mathrm{B}^{a}_{\mathrm{RK},j}(\mathbf{Q}^{n+1}, \mathbf{A}^{n+1}) &\equiv \mathcal{R}^{\mathrm{semi}}_{a}(w_j; \mathbf{Q}^{n+1}, \mathbf{A}^{n+1}) = 0 \equiv \mathrm{f}^{a}_{\mathrm{RK},j}. \end{aligned} \tag{B.37}$$

Now, this allows a Newton-like method which has a Jacobian of the form:

$$\frac{\partial(\mathrm{B}^{u}_{\mathrm{RK},j}, \mathrm{B}^{a}_{\mathrm{RK},j})}{\partial(\mathbf{Q}^{n+1}, \mathbf{A}^{n+1})} = \begin{bmatrix} \left(\frac{1}{\Delta t}\frac{\partial \mathcal{R}^{t}_{u}}{\partial \mathbf{Q}}\right)\Big|_{w=w_j} & \left(\frac{1}{\Delta t}\frac{\partial \mathcal{R}^{t}_{u}}{\partial \mathbf{A}}\right)\Big|_{w=w_j} \\ \frac{\partial \mathcal{R}^{\mathrm{semi}}_{a}}{\partial \mathbf{Q}}\Big|_{w=w_j} & \frac{\partial \mathcal{R}^{\mathrm{semi}}_{a}}{\partial \mathbf{A}}\Big|_{w=w_j} \end{bmatrix}, \tag{B.38}$$

Now, this results in a system that should be nonsingular whenever $\mathcal{R}_a^{\text{semi}}$ is nonsingular, which is broadly the case. Thus, we can advance the state and auxiliary completely to t^{n+1}!

C

Cost model validation for DGBR4/DIRK-DAE solutions of Kuramoto-Sivashinsky equation

In order to assess the wallclock costs of solutions of the DGBR4/DIRK-DAE scheme used in Chapter 3 and developed in Appendix B, we assess the cost model in this appendix.

Consider the solution of the Kuramoto-Sivashinsky equation using DGBR4 with DIRK-DAE timestepping. In Section 3.3, we establish a cost model for s-stage diagonally implicit Runge-Kutta (DIRK) scheme using an order p_x DGBR4 scheme for the spatial solution:

$$\begin{aligned}\mathcal{C}_t &= \underbrace{\left[(p_x+1)^3 N_{\text{elem}}\right]}_{\text{cost of linear solve}} \underbrace{\left[(s+1)N_t\right]}_{\text{timestepping cost}} \qquad \text{(3.45, reprise)}\\ &= \mathcal{C}_{p_x s} N_{\text{elem}} N_t,\end{aligned}$$

where

$$\mathcal{C}_{p_x s} = (p_x+1)^3(s+1). \qquad \text{(3.46, reprise)}$$

In Chapter 3, these are derived under the assumption $s = p_t$.

In this Appendix, we demonstrate that this model adequately describes the wallclock time. We have run the KSE using the DGBR4/DIRK-DAE scheme in this section and performed a set of experiments with $N_t = 100$ timesteps each and using $s = 3$ (RK3) for demonstration. In a first experiment, we simulate with $L = 128.0$, $t_0 = 0$, and $T_s = 1.0$ and show that the wallclock time to solution scales with $(p_x+1)^3 N_{\text{elem}}$. The results of this study are shown in Figure C.1. In this plot, we can see that for each of $p_x = 1$, $p_x = 2$, and $p_x = 3$, the

Simulations in this study are performed on one machine, in one session, and with isolated computational load. For that reason, timing results from this section should be understood to be arbitrarily united, but consistent.

the scaling of wallclock time is linear with respect to the linear cost model, as we expect.

In addition to the spatial model of this form, we also assume that there will not be variation of the cost of solution of the nonlinear system as a function of Δt, due to increased numbers of nonlinear iterations due to system stiffness. In Figure C.2, we study the dependence of wallclock time on Δt with $L = 128$, $N_{\text{elem}} = 128$, $t_0 = 0$, $T_s = 1.0$, and $s = 3$. While there is some stiffness effect for large Δt, the wallclock time to solution for simulations subject to this effect remain within a factor of 1.5 of the small-Δt asymptotic value for timestep sizes up to $\Delta t = 1$. Given this bound on these effects, we conclude that the error model in (3.45) will have satisfactory performance as a model for wallclock time to solution.

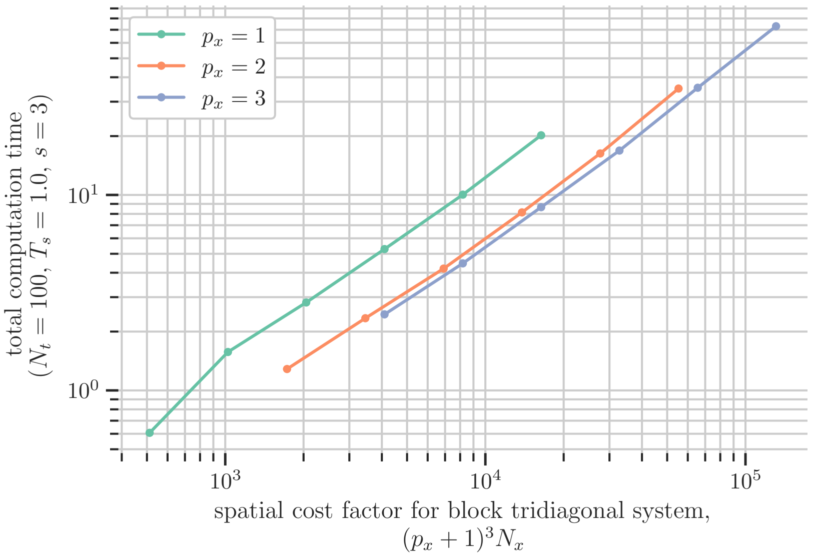


Figure C.1: Wallclock time study at fixed T_s, $N_t = 100$ as a function of the modeling terms for the linear system solve cost.

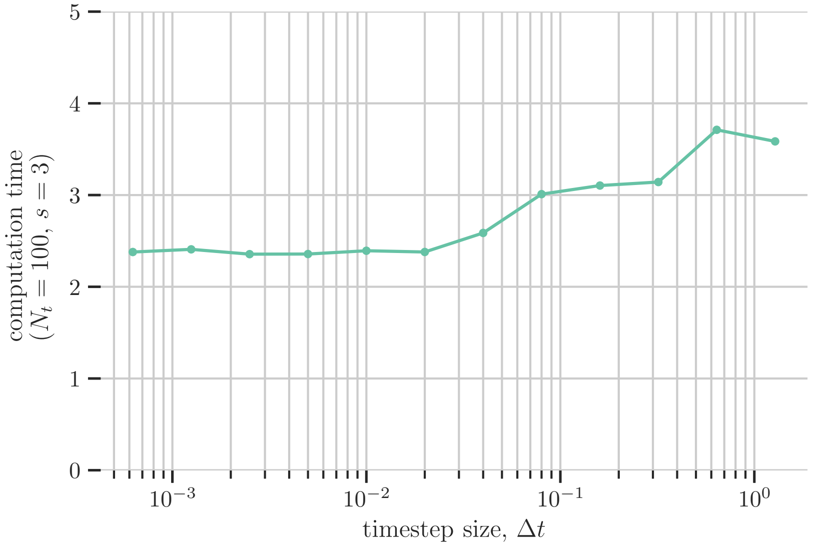


Figure C.2: Wallclock time study at fixed $t_0 = 0$, $T_s = 1.0$, $N_{\text{elem}} = 128$ as a function of the Δt.

D

Bayesian fit testing: synthetic test data

In Section 3.6, we developed a small-sample estimator that could identify the discretization and sampling error behavior nominally on the attractor of a chaotic PDE problem. In order to determine the source of inaccuracies in the small-sample process demonstrated in Chapter 3, we will now assess the performance of the small-sample estimators on synthetic data, generated using the asymptotic theoretical behavior:

$$J_{T,hp}^{\mathrm{syn}} \sim \mathcal{N}\left(J_\infty + C_{p_x}^* \Delta x^{q_x} + B_{p_t}^* \Delta t^{q_t}, \left(A_0^* T_s^{-r}\right)^2\right), \tag{D.1}$$

where the values from Table 3.2 and (3.47) for $p_x = 1$ and $p_t = 2$ are used.

At a limit of $\mathcal{C} = 10^{10}$ and using the sampling costs of $\mathcal{C}_s = 10^7$, a set of 100,000 synthetic output datapoints, $\{J_{T,hp}^{\mathrm{syn}}\}$, are generated, the errors in which can be seen in Figure D.1.

For each forthcoming Bayesian fit, we will draw a length-M set $\{J_{T,hp}\}$ of simulation results using a random number generator. Then, $\{J_{T,hp}\}$, (3.35), (3.36), and (3.41) are used to compute the MAP estimate θ_{MAP}^* as well as an estimate of θ_{MAP} and the error models using (3.33). In Figure D.2, the first such Bayesian fit is shown, with $M = 1000$ points.

Quantitatively, the fit has a small mismatch with the true value of the optimal error, and a slight offset of the optimizer. The values and their comparison to the values in Table 3.2 are given in Table D.1.

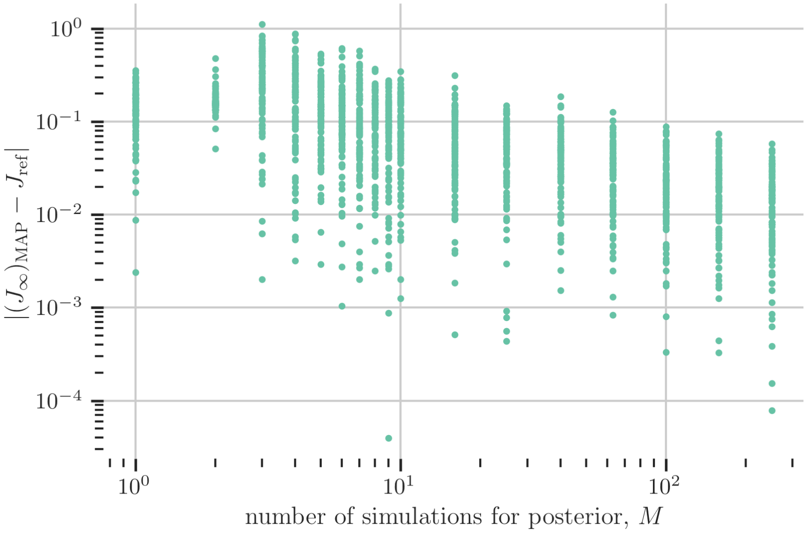


Figure D.1: Error values of synthetic output error data points for $\mathcal{C}_s = 10^7$ and $t_0 = 12{,}000$. 10,000 total points.

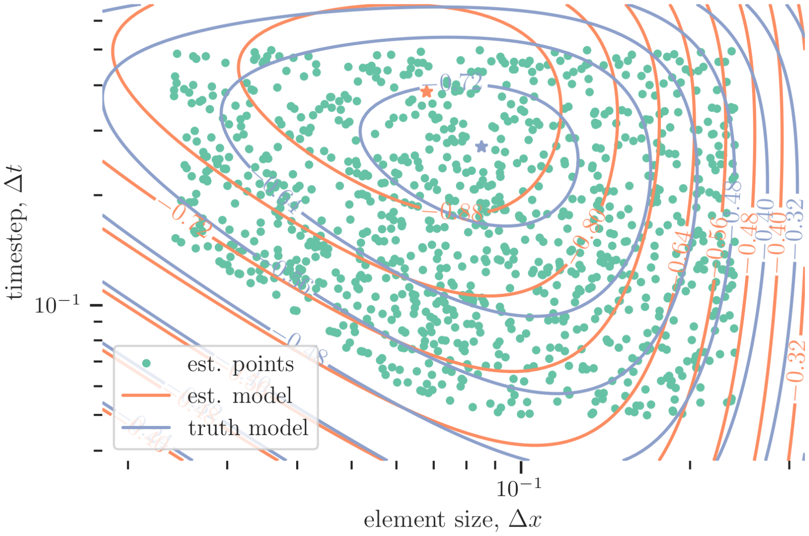


Figure D.2: Bayesian error model fit with $M = 1000$ synthetic datapoints drawn randomly. Small-sample estimate shown with white contours, "true" model shown with black and colored background contours.

These results show that the error model captures J_∞, $C^*_{p_x}$, and A^*_0 under 5% error, while the temporal discretization error models have larger errors. To further interrogate these results, we can look at the convergence behavior as understood by the model in Figure D.3.

variable	value	pct. error
$(C^*_{p_x})_{\mathrm{MAP}}$	502.63	4.62%
$(B^*_{p_t})_{\mathrm{MAP}}$	−47.60	0.12%
$(A^*_0)_{\mathrm{MAP}}$	69.79	2.48%
$(J_\infty)_{\mathrm{MAP}}$	118.53	0.07%
variable	value	ref. value
$(\Delta x_{\mathrm{opt}})_{\mathrm{MAP}}$	0.083	0.085
$(\Delta t_{\mathrm{opt}})_{\mathrm{MAP}}$	0.269	0.271
$(e_{\mathrm{opt}})_{\mathrm{MAP}}$	20.733	20.921

Table D.1: Small-sample fit results for $p_x = 1$, RK2 with $M = 1000$.

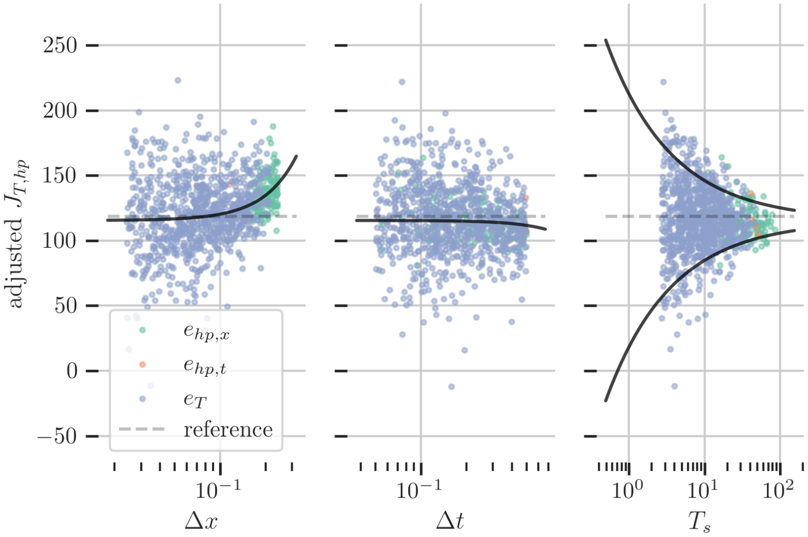


Figure D.3: Convergence behavior under small-sample fit for $p_x = 1$, RK2 with $M = 1000$. Adjusted $J_{T,hp}$ removes modeled temporal effects for spatial convergence, vice versa, and removes both spatial and temporal effects for the sampling error convergence. Data colored by dominant effect under observed model.

These results show that temporal-discretization-dominated simulations are relatively few, compared to the spatial-discretization- and sampling-dominated simulations.

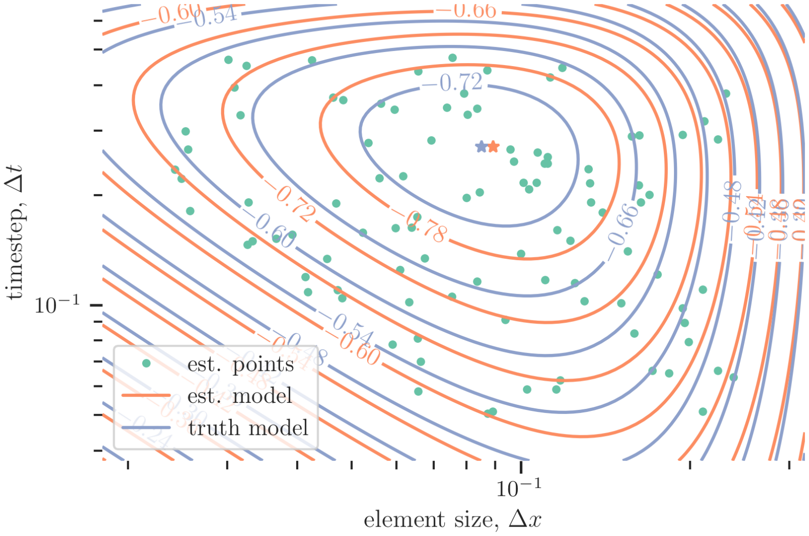


Figure D.4: Bayesian error model fit with $M = 100$ synthetic datapoints drawn randomly.

variable	value	pct. error
$(C^*_{p_x})_{\mathrm{MAP}}$	479.51	0.19%
$(B^*_{p_t})_{\mathrm{MAP}}$	−43.33	8.86%
$(A^*_0)_{\mathrm{MAP}}$	55.65	22.25%
$(J_\infty)_{\mathrm{MAP}}$	119.88	1.22%
variable	value	ref. value
$(\Delta x_{\mathrm{opt}})_{\mathrm{MAP}}$	0.061	0.066
$(\Delta t_{\mathrm{opt}})_{\mathrm{MAP}}$	0.201	0.211
$(e_{\mathrm{opt}})_{\mathrm{MAP}}$	10.542	12.665

Table D.2: Small-sample fit results for $p_x = 1$, RK2 with $M = 100$.

In Figures D.4 and D.5, the fit results for $M = 100$ and $M = 10$ simulations are shown, respectively, with the corresponding data in Tables D.2 and D.3.

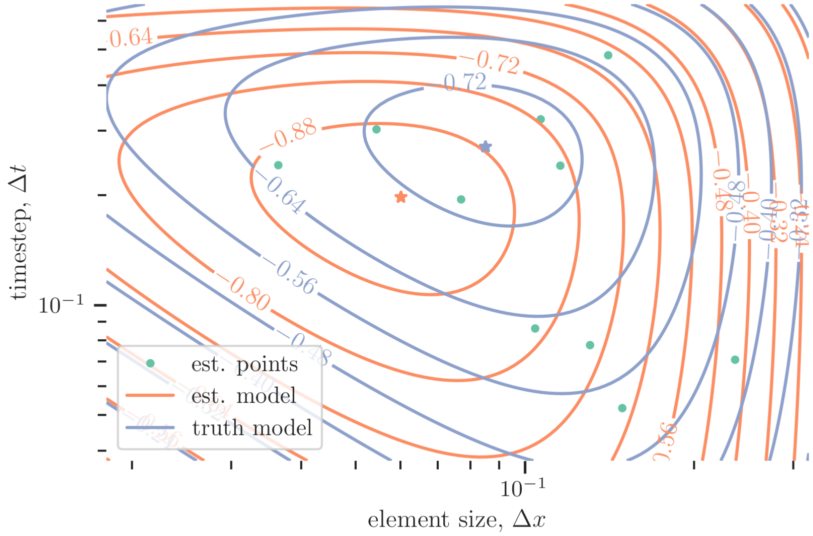


Figure D.5: Bayesian error model fit with $M = 10$ synthetic datapoints drawn randomly. Small-sample estimate shown with white contours, "true" model shown with black and colored background contours.

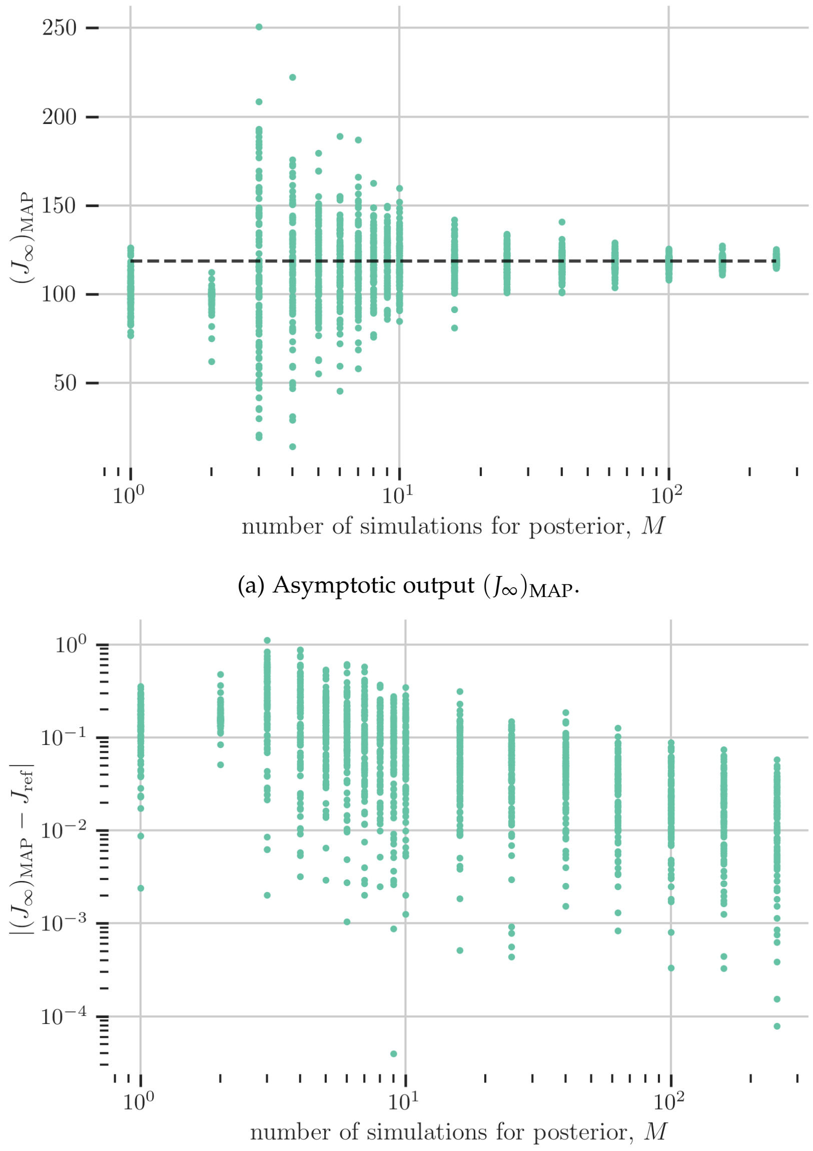
These anecdotal results for $M = 1000$, 100, and 10 demonstrate that the small-sample procedure can identify J_∞ and make high-quality qualitative estimates of the optimal cost-constrained discretization $(\Delta x_{\mathrm{opt}}, \Delta t_{\mathrm{opt}})$, even approaching the small-sample limit using synthetic data.

variable	value	pct. error
$(C^*_{p_x})_{\mathrm{MAP}}$	465.70	3.06%
$(B^*_{p_t})_{\mathrm{MAP}}$	−41.72	12.24%
$(A^*_0)_{\mathrm{MAP}}$	36.20	49.42%
$(J_\infty)_{\mathrm{MAP}}$	118.59	0.12%
variable	value	ref. value
$(\Delta x_{\mathrm{opt}})_{\mathrm{MAP}}$	0.053	0.066
$(\Delta t_{\mathrm{opt}})_{\mathrm{MAP}}$	0.177	0.211
$(e_{\mathrm{opt}})_{\mathrm{MAP}}$	7.826	12.665

Table D.3: Small-sample fit results for $p_x = 1$, RK2 with $M = 10$.

Finally, we study the sensitivity of the small-sample procedure to the particular set of M simulations of $J_{T,hp}$. In order to do so, we will now repeat the process of the small-sample fits over $\mathcal{M} = 100$ length-M sets. Each of these sets will be generated by random number generator draws from the synethetic data likelihood function. Due to computational limits, $\mathcal{M} = 100$ is not available for every choice of M.

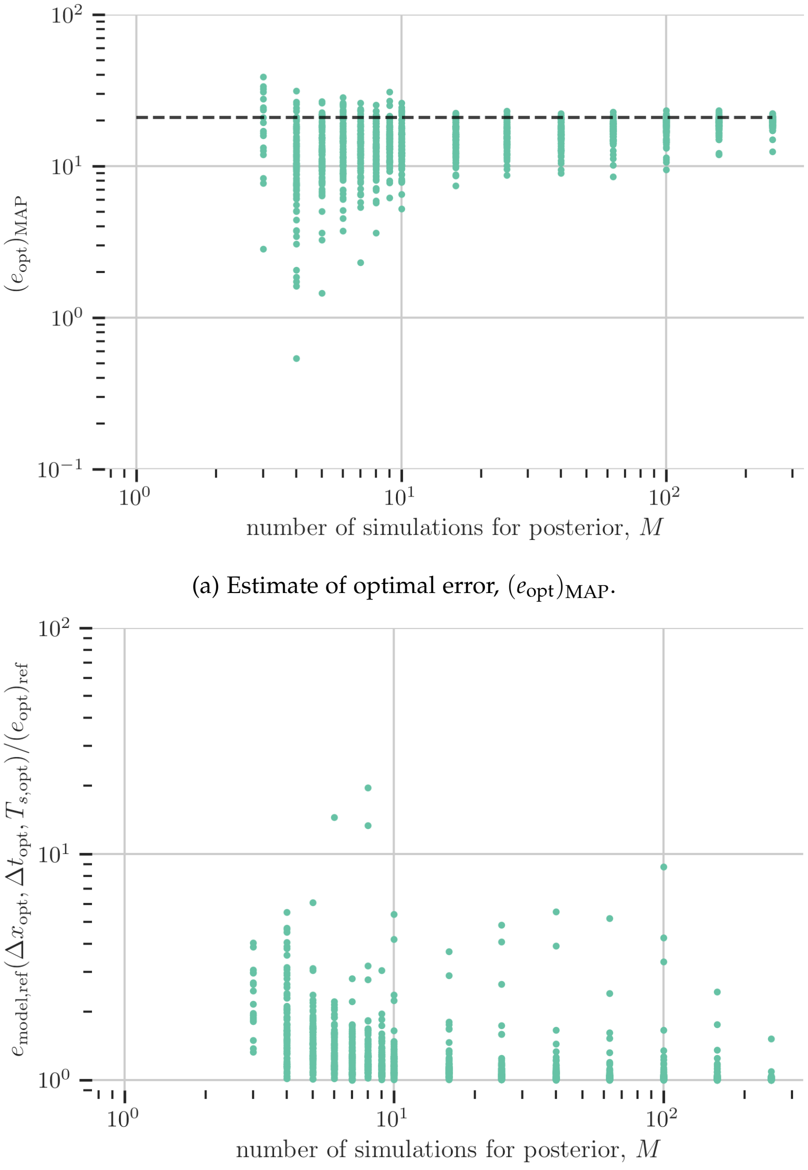
In Figure D.6, the $(J_\infty)_{\mathrm{MAP}}$ estimates are shown. This plot demonstrates that as M increases, the distribution of $(J_\infty)_{\mathrm{MAP}}$ converges towards J_{ref} with a rate approximately $M^{-1/2}$. In Figure D.7, we consider the accuracy of error estimates and optimized discretizations that are generated by each of the length-M MAP estimates. In these plots, we see the quality of the error model, measured by the approach of $(e_{\mathrm{opt}})_{\mathrm{MAP}}$ to $(e_{\mathrm{opt}})_{\mathrm{ref}}$, is good and stays within a small factor of the reference error. In the lower subplot of Figure D.7, we show the multiplicative factor between the error under the reference model at $(e_{\text{opt}})_{\text{MAP}}$ and the optimal error using the reference model, which should represent the lowest possible error at this cost. This plot shows that, using the small sample estimation capability here, more than 90% of the estimates with random length-M sets $\{J_{T,hp}\}$ with $M \geq 10$ should result in a simulation with no more than a factor of two more than the minimum possible error.



(a) Asymptotic output $(J_\infty)_{\mathrm{MAP}}$.

(b) Error $|(J_\infty)_{\mathrm{MAP}} - J_{\mathrm{ref}}|$.

Figure D.6: Small sample identification of J_∞ as a function of M.



(a) Estimate of optimal error, $(e_{\mathrm{opt}})_{\mathrm{MAP}}$.

(b) Excess error in optimized $(M+1)$-th simulation.

Figure D.7: Small sample error estimation as a function of M. In (a): 288 total points out of range: 100 at $M = 1$, 100 at $M = 2$, 88 at $M = 3$. In (b): 298 total points out of range: 100 at $M = 1$, 100 at $M = 2$, 88 at $M = 3$, 3 at $M = 4$, 1 at $M = 6$, 1 at $M = 8$, 1 at $M = 9$, 2 at $M = 10$, 1 at $M = 63$, 1 at $M = 158$.

Bibliography

Gene M. Amdahl. Validity of the single processor approach to achieving large scale computing capabilities. In *Proceedings of the April 18-20, 1967, Spring Joint Computer Conference*, AFIPS '67 (Spring), page 483–485, New York, NY, USA, 1967. Association for Computing Machinery.

V. Araújo, I. Melbourne, and P. Varandas. Rapid mixing for the Lorenz attractor and statistical limit laws for their time-1 maps. *Communications in Mathematical Physics*, 340(3):901–938, 2015.

F. Bassi and S. Rebay. A high-order accurate discontinuous finite element method for the numerical solution of the compressible Navier-Stokes equations. *Journal of Computational Physics*, 131(2): 267–279, 1997.

Roland Becker and Rolf Rannacher. An optimal control approach to a posteriori error estimation in finite element methods. *Acta Numerica*, 10:1–102, 2001.

Patrick J. Blonigan and Qiqi Wang. Least squares shadowing sensitivity analysis of a modified Kuramoto–Sivashinsky equation. *Chaos, Solitons & Fractals*, 64:16–25, 2014.

Patrick J. Blonigan, Pablo Fernandez, Scott M. Murman, Qiqi Wang, Georgios Rigas, and Luca Magri. Toward a chaotic adjoint for LES. In *Proceedings of the Summer Program*. Center for Turbulence Research, Stanford University, 2016.

J. Bodart and J. Larsson. Sensor-based computation of transitional flows using wall-modeled large eddy simulation. In *Annual Research Briefs*, pages 229–240. Center for Turbulence Research, Stanford University, 2012.

Sanjeeb T. Bose and George Ilhwan Park. Wall-modeled large-eddy simulation for complex turbulent flows. *Annual Review of Fluid Mechanics*, 50(1):535–561, 2018.

Richard C. Bradley. Basic properties of strong mixing conditions. a survey and some open questions. *Probability Surveys*, 2:107 – 144, 2005.

Hugh A. Carson, Arthur C. Huang, Marshall C. Galbraith, Steven R. Allmaras, and David L. Darmofal. Mesh Optimization via Error Sampling and Synthesis: An update. In *AIAA Scitech Forum*, 2020.

Nisha Chandramoorthy and Qiqi Wang. Sensitivity computation of statistically stationary quantities in turbulent flows. In *AIAA Aviation Forum*, 2019.

Nisha Chandramoorthy and Qiqi Wang. On the probability of finding nonphysical solutions through shadowing. *Journal of Computational Physics*, 440, 2021.

Dean R. Chapman. Computational aerodynamics development and outlook. *AIAA Journal*, 17(12):1293–1313, 1979.

Haecheon Choi and Parviz Moin. Grid-point requirements for large eddy simulation: Chapman's estimates revisited. *Physics of Fluids*, 24(1):011702, 2012.

Philippe G. Ciarlet and Pierre-Arnaud Raviart. A mixed finite element method for the biharmonic equation. In Carl de Boor, editor, *Mathematical Aspects of Finite Elements in Partial Differential Equations*, pages 125–145. Academic Press, 1974.

Bernardo Cockburn. *An introduction to the Discontinuous Galerkin method for convection-dominated problems*, pages 150–268. Springer Berlin Heidelberg, Berlin, Heidelberg, 1998. Lectures given at the 2nd Session of the Centro Internazionale Matematico Estivo (C.I.M.E.) held in Cetraro, Italy, June 23–28, 1997.

Bernardo Cockburn, George E. Karniadakis, and Chi-Wang Shu. The development of discontinuous Galerkin methods. In Bernardo Cockburn, George E. Karniadakis, and Chi-Wang Shu, editors, *Discontinuous Galerkin Methods*, pages 3–50. Springer Berlin Heidelberg, Berlin, Heidelberg, 2000.

Corentin Carton de Wiart and Scott M. Murman. Assessment of wall-modeled LES strategies within a discontinuous-Galerkin spectral-element framework. In *55th AIAA Aerospace Sciences Meeting*.

Juan C. Del Álamo, Javier Jiménez, Paulo Zandonade, and Robert D. Moser. Scaling of the energy spectra of turbulent channels. *Journal of Fluid Mechanics*, 500:135–144, 2004.

Manfred Denker. The central limit theorem for dynamical systems. *Dynamical Systems and Ergodic Theory*, 23:33–62, 1989.

Laslo T. Diosady and Scott M. Murman. Higher-order methods for compressible turbulent flows using entropy variables. In *Proceedings of the 53rd AIAA Aerospace Sciences Meeting*, 2015.

Michael G. Dodson and David S. Miklosovic. An historical and applied aerodynamic study of the Wright Brothers' wind tunnel test program and application to successful manned flight. In *Fluids Engineering Division Summer Meeting*, volume 1, pages 269–278. American Society of Mechanical Engineers, 06 2005. Symposia, Parts A and B.

J. R. Dormand, R. R. Duckers, and P. J. Prince. Global error estimation with Runge-Kutta methods. *IMA Journal of Numerical Analysis*, 4(2): 169–184, 04 1984.

J.-P. Eckmann and D. Ruelle. Ergodic theory of chaos and strange attractors. In *The Theory of Chaotic Attractors*, pages 273–312. Springer, New York, NY, 2004.

Pablo Fernandez and Qiqi Wang. Lyapunov spectrum of the separated flow around the NACA 0012 airfoil and its dependence on numerical discretization. *Journal of Computational Physics*, 350: 453–469, 2017.

Benedetta Ferrario. Ergodic results for stochastic Navier-Stokes equation. *Stochastics*, 60(3-4):271–288, 1997.

Franco Flandoli and Bohdan Maslowski. Ergodicity of the 2-D Navier-Stokes equation under random perturbations. *Communications in Mathematical Physics*, 172(1):119–141, 1995.

David A. Freedman. On the asymptotic behavior of Bayes' estimates in the discrete case. *The Annals of Mathematical Statistics*, 34(4): 1386–1403, 12 1963.

Emmanuil H. Georgoulis, Paul Houston, and Juha Virtanen. An a posteriori error indicator for discontinuous Galerkin approximations of fourth-order elliptic problems. *IMA Journal of Numerical Analysis*, 31(1):281–298, 09 2009.

Gianluca Geraci, Michael S. Eldred, and Gianluca Iaccarino. A multifidelity multilevel Monte Carlo method for uncertainty propagation in aerospace applications. In *19th AIAA Non-Deterministic Approaches Conference*, 2017.

F. Ginelli, P. Poggi, A. Turchi, H. Chaté, R. Livi, and A. Politi. Characterizing dynamics with Covariant Lyapunov Vectors. *Phys. Rev. Lett.*, 99:130601, Sep 2007.

Konrad A. Goc, Oriol Lehmkuhl, George Ilhwan Park, Sanjeeb T. Bose, and Parviz Moin. Large eddy simulation of aircraft at affordable cost: a milestone in computational fluid dynamics. *Flow*, 1, 2021.

C. Gorlé and G. Iaccarino. A framework for epistemic uncertainty quantification of turbulent scalar flux models for Reynolds-averaged Navier-Stokes simulations. *Physics of Fluids*, 25(5):055105, 2013.

John L. Gustafson. Reevaluating Amdahl's law. *Commun. ACM*, 31(5): 532–533, May 1988.

Ernst Hairer and Gerhard Wanner. Number 14 in Springer Series in Computational Mathematics. Springer, Berlin, Heidelberg, second edition, 1993.

Martin Hairer and Jonathan C. Mattingly. Ergodicity of the 2d Navier-Stokes equations with degenerate stochastic forcing. *Annals of Mathematics*, 164(3):993–1032, 2006.

Jack K. Hale and Geneviéve Raugel. Upper semicontinuity of the attractor for a singularly perturbed hyperbolic equation. *Journal of Differential Equations*, 73(2):197–214, 1988.

F.J. Harris. On the use of windows for harmonic analysis with the discrete Fourier transform. *Proceedings of the IEEE*, 66(1):51–83, 1978.

Mark D. Hill and Michael R. Marty. Amdahl's Law in the multicore era. *Computer*, 41(7):33–38, 2008.

Xun Huan and Youssef M. Marzouk. Simulation-based optimal Bayesian experimental design for nonlinear systems. *Journal of Computational Physics*, 232(1):288–317, 2013.

Arthur Chan-wei Huang. *An adaptive variational multiscale method with discontinuous subscales for aerodyanamic flows*. PhD thesis, Massachusetts Institute of Technology, 2020.

Thomas J.R. Hughes and Gregory M. Hulbert. Space-time finite element methods for elastodynamics: Formulations and error estimates. *Computer Methods in Applied Mechanics and Engineering*, 66(3):339–363, 1988.

Yashod Savithru Jayasinghe. *An adaptive space-time discontinuous Galerkin method for reservoir flows*. PhD thesis, Massachusetts Institute of Technology, 2018.

Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1):99–134, 1998.

George Karniadakis and Spencer Sherwin. *Spectral/hp element methods for computational fluid dynamics*. Oxford University Press, 06 2005.

John Kim, Parviz Moin, and Robert Moser. Turbulence statistics in fully developed channel flow at low Reynolds number. *Journal of Fluid Mechanics*, 177:133–166, 1987. DOI: 10.1017/S0022112087000892.

Peter E. Kloeden and Jens Lorenz. Stable attracting sets in dynamical systems and in their one-step discretizations. *SIAM Journal on Numerical Analysis*, 23:986–995, 1986.

Yoshiki Kuramoto. Diffusion-induced chaos in reaction systems. *Progress of Theoretical Physics Supplement*, 64:346–367, 02 1978.

Johan Larsson. Grid-adaptation for chaotic multi-scale simulations as a verification-driven inverse problem. In *AIAA Aerospace Sciences Meeting*, 2018.

Daniel J. Lea, Myles R. Allen, and Thomas W.N. Haine. Sensitivity analysis of the climate of a chaotic system. *Tellus A: Dynamic Meteorology and Oceanography*, 52(5):523–532, 2000.

Myoungkyu Lee and Robert D. Moser. Direct numerical simulation of turbulent channel flow up to $\mathrm{Re}_\tau \approx 5200$. *Journal of Fluid Mechanics*, 774:395–415, 2015.

David Levy, Kelly Laflin, John Vassberg, Edward Tinoco, Mortaza Mani, Ben Rider, Olaf Brodersen, Simone Crippa, Christopher Rumsey, Richard Wahls, Joe Morrison, Dimitri Mavriplis, and Mitsuhiro Murayama. Summary of data from the fifth AIAA CFD Drag Prediction Workshop. In *51st AIAA Aerospace Sciences Meeting*, 2013.

Michael James Lighthill, John Michael Tutill Thompson, A. K. Sen, A. G. M. Last, D. J. Tritton, Basil John Mason, P. Mathias, and John Hugh Westcott. The recently recognized failure of predictability in Newtonian dynamics. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, 407(1832): 35–50, 1986.

Edward N. Lorenz. Deterministic nonperiodic flow. *Journal of Atmospheric Sciences*, 20(2):130 – 141, 1963.

Adrián Lozano-Durán and Javier Jiménez. Effect of the computational domain on direct simulations of turbulent channels up to $\mathrm{Re}_\tau = 4200$. *Physics of Fluids*, 26(1):011702, 2014.

Chris A Mack. Fifty years of Moore's law. *IEEE Transactions on Semiconductor Manufacturing*, 24(2):202–207, 2011.

Vakhtang Makarashvili, Elia Merzari, Aleksandr Obabko, Andrew Siegel, and Paul Fischer. A performance analysis of ensemble averaging for high fidelity turbulence simulations at the strong scaling limit. *Computer Physics Communications*, 219:236–245, 2017.

Jonathan C. Mattingly. Ergodicity of 2D Navier-Stokes equations with random forcing and large viscosity. *Communications in Mathematical Physics*, 206(2):273–288, 1999.

Charles Mockett, Thilo Knacke, and Frank 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*, pages 9–11. European Research Collaboration on Flow Turbulence and Combustion, 2010.

Parviz Moin and Krishnan Mahesh. Direct numerical simulations: A tool in turbulence research. *Annual Review of Fluid Mechanics*, 30(1): 539–578, 1998.

Florian Müller, Patrick Jenny, and Daniel W. Meyer. Parallel multilevel Monte Carlo for two-phase flow and transport in random heterogeneous porous media with sampling-error and discretization-error balancing. *SPE Journal*, 21(06):2027–2037, 09 2016.

Daniel J. Navarro, Ben R. Newell, and Christin Schulze. Learning and choosing in an uncertain world: An investigation of the explore-exploit dilemma in static and dynamic environments. *Cognitive Psychology*, 85:43–77, 2016.

Todd A. Oliver, Nicholas Malaya, Rhys Ulerich, and Robert D. Moser. Estimating uncertainties in statistics computed from direct numerical simulation. *Physics of Fluids*, 26(3):035101, 2014.

George Ilhwan Park and Parviz Moin. An improved dynamic non-equilibrium wall-model for large eddy simulation. *Physics of Fluids*, 26(1):015108, 2014.

Benjamin Peherstorfer, Philip S. Beran, and Karen E. Willcox. Multifidelity Monte Carlo estimation for large-scale uncertainty propagation. In *AIAA Non-Deterministic Approaches Conference*, 2018.

Niles A. Pierce and Michael B. Giles. Adjoint recovery of superconvergent functionals from PDE approximations. *SIAM Review*, 42(2):247–264, 2000.

U. Piomelli. Large-eddy simulation: achievements and challenges. *Progress in Aerospace Sciences*, 35(4):335–362, 1999.

Ugo Piomelli and Elias Balaras. Wall-layer models for large-eddy simulations. *Annual Review of Fluid Mechanics*, 34(1):349–374, 2002.

Ashur Rafiev, Mohammed A. N. Al-Hayanni, Fei Xia, Rishad Shafik, Alexander Romanovsky, and Alex Yakovlev. Speedup and power scaling models for heterogeneous many-core systems. *IEEE Transactions on Multi-Scale Computing Systems*, 4(3):436–449, 2018.

David Ruelle. Differentiation of SRB states for hyperbolic flows. *Ergodic Theory and Dynamical Systems*, 28(2):613–631, 2008.

Serena Russo and Paolo Luchini. A fast algorithm for the estimation of statistical error in DNS (or experimental) time averages. *Journal of Computational Physics*, 347:328–340, 2017.

George R. Sell. Global attractors for the three-dimensional Navier-Stokes equations. *Journal of Dynamics and Differential Equations*, 8(1): 1–33, 1996.

Gregory I. Sivashinsky. Nonlinear analysis of hydrodynamic instability in laminar flames– I. derivation of basic equations. *Acta Astronautica*, 4(11):1177–1206, 1977.

Isaac J Sledge and José C Príncipe. Balancing exploration and exploitation in reinforcement learning using a value of information criterion. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 2816–2820. IEEE, 2017.

Jeffrey P Slotnick, Abdollah Khodadoust, Juan Alonso, David Darmofal, William Gropp, Elizabeth Lurie, and Dimitri J Mavriplis. CFD Vision 2030 study: a path to revolutionary computational aerosciences. Technical Report CR–2014-218178, NASA, 2014.

Alexander J. Smits, Beverley J. McKeon, and Ivan Marusic. High–Reynolds number wall turbulence. *Annual Review of Fluid Mechanics*, 43(1):353–375, 2011.

Jonathan Sorg, Satinder Singh, and Richard L. Lewis. Variance-based rewards for approximate Bayesian reinforcement learning. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, UAI'10, page 564–571, Arlington, Virginia, USA, 2010. AUAI Press.

Philippe R. Spalart. Detached-eddy simulation. *Annual Review of Fluid Mechanics*, 41(1):181–202, 2009.

P.R. Spalart, W.H. Jou, M. Strelets, S.R. Allmaras, et al. Comments on the feasibility of LES for wings, and on a hybrid RANS/LES approach. In *Proceedings of the 1st AFOSR Int. Conf. on DNS/LES*, volume 1, pages 4–8, 1997.

Colin Sparrow. *The Lorenz equations*. Applied Mathematical Sciences. Springer, 1982.

Steven H. Strogatz. *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering*. CRC Press, 2 edition, 2015.

Andrew M. Stuart. Numerical analysis of dynamical systems. *Acta Numerica*, 3:467–572, 1994.

Xian-He Sun and Lionel M Ni. Another view on parallel speedup. In *Supercomputing '90: Proceedings of the 1990 ACM/IEEE Conference on Supercomputing*, pages 324–333, 1990.

Stan Developer Team. Stan modeling language users guide and reference manual. Technical report, 2021.

Roney L. Thompson, Luiz Eduardo B. Sampaio, Felipe A.V. de Bragança Alves, Laurent Thais, and Gilmar Mompean. A methodology to evaluate statistical errors in DNS data of plane channel flows. *Computers & Fluids*, 130:1–7, 2016.

Edward N. Tinoco, Olaf P. Brodersen, Stefan Keye, Kelly R. Laflin, Edward Feltrop, John C. Vassberg, Mori Mani, Ben Rider, Richard A. Wahls, Joseph H. Morrison, David Hue, Christopher J. Roy, Dimitri J. Mavriplis, and Mitsuhiro Murayama. Summary data from the sixth AIAA CFD Drag Prediction Workshop: CRM cases. *Journal of Aircraft*, 55(4):1352–1379, 2018.

Kevin E. Trenberth. Some effects of finite sample size and persistence on meteorological statistics. Part I: Autocorrelations. *Monthly Weather Review*, 112(12):2359 – 2368, 1984.

Divakar Viswanath. Global errors of numerical ODE solvers and Lyapunov's theory of stability. *IMA Journal of Numerical Analysis*, 21 (1):387–406, 01 2001.

Qiqi Wang, Rui Hu, and Patrick Blonigan. Least Squares Shadowing sensitivity analysis of chaotic limit cycle oscillations. *Journal of Computational Physics*, 267:210–224, 2014.

Yan Xu and Chi-Wang Shu. Local discontinuous Galerkin methods for the Kuramoto-Sivashinsky equations and the Ito-type coupled KdV equations. *Computer Methods in Applied Mechanics and Engineering*, 195(25):3430–3447, 2006.

Masayuki Yano. *An optimization framework for adaptive higher-order discretizations of partial differential equations on anisotropic simplex meshes*. PhD thesis, Massachusetts Institute of Technology, 2012.

Masayuki Yano and David L. Darmofal. An optimization-based framework for anisotropic simplex mesh adaptation. *Journal of Computational Physics*, 231(22):7626–7649, 2012.