Invited Articles

Advancing understanding of turbulence through extreme-scale computation: Intermittency and simulations at large problem sizes

P. K. Yeung *
School of Aerospace Engineering and School of Mechanical Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, USA

K. Ravikumar
School of Aerospace Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, USA

(Received 5 July 2020; accepted 16 October 2020; published 24 November 2020)

Sustained and rapid advances in computing have enabled the conduct of direct numerical simulations (DNS) at increasing problem sizes and higher levels of physical realism, which have in turn contributed to many advances in understanding turbulence. However, continuing and future success at the "extreme-scale" level will likely require new algorithms adapted to emerging heterogeneous architectures, and even then, long simulations at extreme problem sizes are probably still too costly. In this paper we first describe the essential elements of an asynchronous parallel algorithm for DNS of incompressible isotropic turbulence, which scales effectively up to $18\,432^3$ resolution (more than 6 trillion grid points) on a world-class IBM-NVIDIA CPU-GPU machine. We then propose a simulation paradigm, built on the idea that, for physical quantities of short timescales, sampling over well-separated snapshots in a long simulation at high resolution can be replaced by sampling over short simulation segments with a high degree of statistical independence evolved from snapshots at modest or even low resolution. The total computational cost is now counted in Kolmogorov timescales instead of large-eddy timescales, leading to tremendous savings at high Reynolds number. This "Multiple Resolution Independent Simulations" (MRIS) approach is validated through a series of comparisons, and subsequently applied to obtain results on fine-scale intermittency, at Taylor-scale Reynolds numbers 390 to 1300, with grid spacing smaller than the Kolmogorov length scale. The results show conclusively that extreme fluctuations of the dissipation rate are usually accompanied by extreme enstrophy, while extreme enstrophy is usually accompanied by less-intense dissipation. Statistics of the locally averaged dissipation and enstrophy suggest these two variables scale together at inertial-range scale sizes (but not in the dissipation range). Finally, brief remarks are made concerning perspectives on likely major challenges in an exascale future, and several other topics of study where the MRIS approach may be useful.

DOI: 10.1103/PhysRevFluids.5.110517

I. INTRODUCTION

In turbulence, it is well known that direct numerical simulations (DNS) at massive scales are very useful for advancing physical understanding, but also very demanding in computational resources [1,2]. For a given flow geometry, strong motivation for ever-larger simulations may include reasons of a physical nature, such as a higher Reynolds number [3], lower diffusivity in turbulent mixing [4], increasing chemical complexity in reacting flows [5], and higher turbulent Mach

*pk.yeung@ae.gatech.edu

numbers in compressible turbulence [6] as well as numerical reasons associated with resolution or sampling in space or time [7,8]. Rapid advances in computing power (exponential growth of roughly 1 million-fold increase over the last 25 years) have enabled simulations of order 1 trillion grid points in at least isotropic turbulence [3], channel flow [9], and stratified turbulence [10]. With exascale computers capable of 10^{18} floating point operations per second (flops) expected to arrive by 2021 or 2022, future prospects for turbulence simulations appear to be bright. However, changes in the fast-evolving high-performance computing (HPC) landscape pose considerable challenges in algorithmic developments necessary to ensure good code performance. Furthermore, turbulence computations are often so demanding that extreme-scale simulations of the largest size that can fit into the computer memory are likely to be restricted to ever-shorter physical time spans.

In this paper we describe some recent progress made, and perspectives developed, in addressing the two challenges indicated above. Of course issues and requirements can vary with the specific flow configuration considered. Here we will limit our attention to isotropic turbulence amenable to the use of Fourier pseudospectral methods on a three-dimensional (3D) domain with N^3 grid points, although similar considerations concerning code performance should hold for homogeneous turbulence in noncubic domains as well. We present simulation results at up to $18\,432^3$ resolution (over 6 trillion grid points), which should help us better understand some of the long-unresolved aspects of intermittency in turbulence [11,12].

Most large parallel fluid dynamics codes employ the distributed-memory programming model, where the solution domain is divided among a (potentially large) number of parallel processes, each operating as a single CPU on its own share of data. However, these processes must share data with each other through calls to standard communication library routines, which adds substantial overhead to the overall cost of the computation. Additional costs also arise from local data movements needed on each parallel process to pack data from different areas of the computer memory into contiguous "messages" before each communication call, and to unpack newly received messages in reverse accordingly. Usually, as the problem size and process count both increase, communication overhead increases and the code becomes less efficient. While most successful large simulations have special measures incorporated to help manage communication costs, and some machines may offer impressively fast communication performance, this is still a major bottleneck (followed closely by local data movements as noted above) in the use of massive CPU-based parallelism.

The last decade in supercomputing has seen the increasing prominence of heterogeneous architectures with tightly coupled large-memory CPUs and smaller-memory GPUs, or other types of hardware accelerators capable of very high computational speed. Because of considerations for energy efficiency, such prominence is likely to continue in the future, even though effective use of such systems is often more challenging [13–15]. Essentially, a new mode of parallelism is now available inside the GPU (which can hold hundreds of execution threads), but data movement necessary between the CPU and the GPU can pose a new challenge, and a substantial increase in algorithmic complexity may be inevitable. Furthermore, it may not be immediately clear to what extent hardware with special strength in faster arithmetic can be of benefit to a code where communication and local data movements (which may occur on CPUs, GPUs, or between them) are inherently dominant. However, it is now possible for operations involving the CPU, GPU, and the data transfer channel between them to occur asynchronously. In Sec. II we discuss the principles of a batched asynchronous algorithm [16] that allows good speedup at extreme problem sizes, as demonstrated on the 200 petaflops supercomputer located at the Oak Ridge National Laboratory, USA, called Summit, which was (according to the URL https://top500.org/lists/top500/2020/06) the world's second fastest as of June 2020.

The quality of results from any DNS depends on the numerical methods used, the degree to which both large scales and small scales are faithfully represented, in both time and space, as well as the adequacy of statistical sampling. For flows with a stationary state, the conventional approach to ensure good sampling is by running a long simulation for, say, $O(10)$ eddy-turnover times (T_E, defined to be the ratio of a longitudinal integral length scale to the r.m.s. fluctuation of a velocity component), performing postprocessing on data saved at regular time intervals, and

finally averaging over multiple realizations. However, since numerical stability (in Runge-Kutta methods and other schemes using explicit discretization in time) requires that the time step Δt scales with the grid spacing (Δx), computational cost for a given physical time period increases at least as fast as N^4. This means every halving of Δx leads to at least a 16 times increase in cost—which exceeds considerably the performance increase available in most newly installed top-ranked machines over their predecessors. It is, in fact, not surprising that most simulations considered leading edge in scale in their time (e.g., Refs. [17,18]) have been relatively short. It should be noted that increasing problem size is in fact being enabled by increases in memory available on leading-edge machines, but actual computational power is increasing more slowly, such that the largest possible simulations performed within finite resource constraints are at risk of being limited to short physical time periods. This leads to the ironic situation that, as computing power grows and algorithms successfully scale to larger problem sizes, the ability to conduct the next-largest long simulations actually becomes increasingly compromised.

In this paper, we introduce a paradigm to address the challenge posed above, for studies of small-scale processes that evolve on short timescales. We first make two observations, which are supported by experience. The first is that statistical stationarity in time, with a mild assumption of ergodicity, allows us to take samples from multiple short *simulation segments*, provided they are well separated in time, with a high degree of statistical independence. The second is that when a modestly resolved velocity field is refined to a higher resolution, the small scales adjust quickly, potentially within a couple of Kolmogorov timescales (τ_η). These observations suggest that an alternative to a long, high-resolution (ideal but unfeasible) simulation may be to start with multiple (say, M) independent lower-resolution snapshots, allow them to quickly adjust to higher resolution, and then start collecting statistics at the highest resolution after only a short period of time (say $\beta\tau_\eta$, with β not much larger than 1). The length of time spent computing on an N^3 grid would then be proportional to $\beta M\tau_\eta$ (in total), as opposed to multiple T_E's. Substantial savings are both most likely while most needed at high Reynolds numbers, where $T_E \gg \tau_\eta$, and it is important [19,20] to resolve down to scales smaller than the Kolmogorov scale ($\eta = (\nu^3/\langle\epsilon\rangle)^{1/4}$) based on the mean energy dissipation rate ($\langle\epsilon\rangle$). We refer to this paradigm as Multiple Resolution Independent Simulations (MRIS). More details, including a validation study, are given in Sec. III.

Some hints to the viability of this approach could be found in recent work [8] where events of extreme dissipation and enstrophy were seen to adjust rapidly to changes in resolution, and important conclusions could be drawn from short simulations of length less than 10 τ_η. In Ref. [8] short simulations of forced isotropic turbulence at two Taylor-scale Reynolds numbers (R_λ, up to 650) were performed at three resolution levels, up to 8192^3 grid points but all starting from the same initial snapshot. This was, in effect, similar to just one MRIS realization of 10 τ_η long. Our current objective is to increase the Reynolds number by running larger simulations, while also performing ensemble averaging over the initial conditions by starting from modestly resolved snapshots originally distributed over several T_E's in time. In addition we study both one- and two-point statistics—in particular, the properties of local averages of the dissipation rate, which play a critical role in understanding intermittency [11,12,21]. Availability of data on such averages over 3D volumes [instead of one-dimensional (1D) versions] is relatively recent [22]. We show some selected results in Sec. IV.

Our intent in this paper is to communicate recent innovations in turbulence simulations that we believe are important to the enduring goal of advancing understanding of turbulence through taking proper advantage of future exascale computing or beyond. The proposed MRIS technique is not all powerful: for example, in our work, a longer adjustment time for the numerical solution will be required when attempting to increase the Reynolds number at a given resolution, than to increase the resolution at a given Reynolds number. A longer adjustment time is likely required also for simulations of wall-bounded turbulence, such as fully developed channel flow spanning several flow-through times [23]. This is in addition to the obvious inapplicability of this approach for simulations with no stationary state. However, this approach is well suited to the task of obtaining well-sampled results of the small-scale physics, at higher resolution in stationary isotropic

turbulence. Sections II– IV are devoted to different aspects of this work, as already indicated in the paragraphs above. Conclusions are summarized in Sec. V.

II. A GPU-OPTIMIZED ALGORITHM FOR PSEUDOSPECTRAL DNS

The basics of Fourier pseudospectral methods [24,25] as well as their implementation on modern parallel computers [26] are well known. We present a few key elements below and then proceed to discuss the structure of a parallel algorithm for extreme problem sizes using GPUs. The latter discussion is meant to highlight some key principles that may be of interest to other turbulence code developers engaged in heterogeneous computing.

A. Computational approach and background

We solve numerically the Navier-Stokes equations for 3D turbulence in the form

$$\partial \mathbf{u}/\partial t = -(\mathbf{u}\cdot\boldsymbol{\nabla})\mathbf{u} - \boldsymbol{\nabla}(p/\rho) + \nu\nabla^2\mathbf{u} + \mathbf{f}, \tag{1}$$

where (assuming constant density) both the fluctuating velocity vector ($\mathbf{u}$) and the forcing term ($\mathbf{f}$) are both solenoidal. In Fourier space we can write

$$\partial \hat{\mathbf{u}}/\partial t = -\{\widehat{\boldsymbol{\nabla}\cdot(\mathbf{uu})}\}_{\perp\mathbf{k}} - \nu k^2\hat{\mathbf{u}} + \hat{\mathbf{f}}, \tag{2}$$

where carets denote Fourier coefficients, $\mathbf{k}$ is the wave-number vector, and $k \equiv |\mathbf{k}|$. In pseudospectral methods the nonlinear dyadic products $\mathbf{uu}$ are first formed in physical space, then transformed to wave-number space and projected onto a plane orthogonal to $\mathbf{k}$. Aliasing errors associated with nonlinear terms are controlled by a combination of phase shifting (for aliasing in one dimension) and truncation at the wave-number magnitude $k_{\max} = \sqrt{2}N/3$ [27,28] for an N^3 grid (which eliminates doubly and triply aliased Fourier modes). Use of the dyadic form as in $\boldsymbol{\nabla}\cdot(\mathbf{uu})$ versus the advective form $(\mathbf{u}\cdot\boldsymbol{\nabla})\mathbf{u}$ has the advantage of reducing the number of variables that need to be Fourier transformed, as well as the level of residual aliasing errors which may arise.

A first decision in our parallel code design is a domain decomposition scheme among P parallel processes, that facilitates the key computational task, i.e.. the 3D fast Fourier transform (FFT), which is taken one direction at a time, on lines of data with all N grid points present in the local memory. The decomposition can be 1D, resulting in $P \leqslant N$ *slabs* each consisting of N/P planes; or two-dimensional (2D), resulting in $P = P_r \times P_c$ *pencils*, each containing $N \times N/P_r \times N/P_c$ grid points. Both schemes have been used in the literature [26,29]. The 2D decomposition allows $P > N$ (in principle, up to N^2) and is well suited for massive parallelism driven by large P. However, a trend towards large-memory multicored processors (*nodes*) and other forms of shared-memory parallelism may favor a 1D decomposition.

B. GPU implementation for extreme problem sizes

Most top-ranked machines possess unique characteristics that explain their ranking. Thus, although portability across different platforms is desirable, writing a code that has significant machine dependency is sometimes a necessary and worthwhile investment. Summit has a heterogeneous architecture, consisting of 4608 nodes, each containing 42 user-addressable IBM Power 9 CPU cores and six NVIDIA Volta GPUs. Each node consists of 512 GB of random access memory (DDR4) for use by the CPU cores and 16 GB of high-bandwidth memory (HBM2) for use by each GPU. Large CPU memory (16 times increase per node compared to the predecessor machine, called Titan) and availability of multiple GPUs per node (versus 1 on Titan) are both important attributes. Clearly, full utilization of all GPUs is easiest if N is divisible by six. Since each node has two sockets, better performance—at the cost of increased programming complexity—is obtained by having two CPU processes tied to three GPUs each, instead of six CPU processes and one GPU each.

Our goal is to reach problem sizes as large as possible, at a level of performance that makes production simulations realistic. The actual FFT computation (as 1D FFTs) using highly optimized

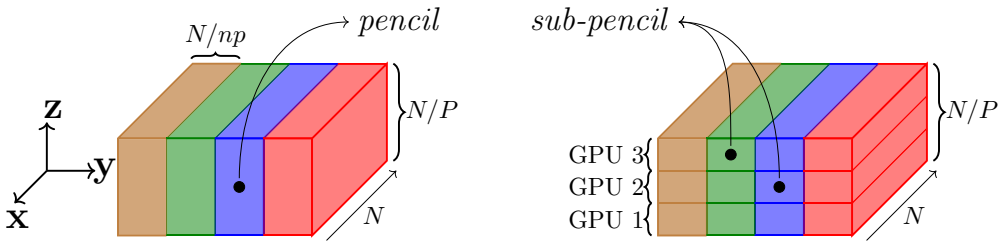

FIG. 1. Left: Decomposition of a *slab* of data into multiple (np) *pencils*, of size $N \times (N/np) \times N/P$ each, for problem sizes where a single *slab* does not fit into the GPU memory. Right: Each *pencil* is further divided to allow batched asynchronism while using multiple GPUs, as done in Ref. [16].

GPU libraries is so fast that its cost is insignificant. The general principles of success then include (1) minimizing communication costs, (2) taking maximum advantage of the CPU and GPU memory, and (3) optimizing the data transfer that must occur. The first of these is much helped by a high network bandwidth on Summit as well as the use of fewer parallel processes. The second involves noting that, although GPU memory is limited, data stored on the large-memory CPU may be divided into smaller units to be processed separately. The third involves using CUDA library calls and specialized GPU *kernels* to transfer data between GPU and CPU while also avoiding any additional costs of packing data into contiguous areas of memory before communication, and the unpacking afterwards.

A relatively simple hybrid CPU-GPU programming model may involve copying an entire slab of data from CPU to the GPU, computing on it and then copying back to the CPU. However, due to the smaller GPU memory this approach would limit the problem sizes accessible. To address GPU memory limitations, we can divide each *slab* along one of its long dimensions, into multiple (np) *pencils* (labeled with distinct colors in Fig. 1). In principle, each GPU can process one pencil as a distinct batch of data. However, if N is very large and np is small, even a single pencil may not fit into the GPU memory, making a further subdivision necessary. A larger np can be used to make each pencil smaller, but this will likely result in each GPU making numerous data transfer calls to cover a larger number of smaller pencils, which is less efficient.

We address the challenges noted above by (assuming N/P is divisible by the number of GPUs per CPU) further subdividing each pencil along its shorter dimension. More importantly, this strategy provides opportunities for asynchronism, or overlapping, that are possible when each GPU works on multiple batches of data drawn from different pencils within the same slab. Each subdivided portion of a pencil is to be processed separately on a GPU: by first being copied in from the CPU, computed on, and finally the results being copied out back to the CPU. This is repeated until all such "subpencils" (within the same slab) assigned to a specific GPU have been processed. For example, as each GPU proceeds from left to right (brown to red in the figure), while computation of data in the blue subpencil proceeds on the GPU, data transfer from GPU to CPU of the (already computed) green subpencil or transfer from CPU to GPU of the red subpencil, which is yet to be computed, can occur simultaneously. Proper scheduling and management of these execution sequences are necessary to ensure correct results is achieved by using CUDA library functions to define two execution *streams*, designated for computation and data transfer respectively. The same protocol of asynchronous operations here is carried out on other GPUs as well, separately from each other. We refer to this strategy as a "batched asynchronism."

With subpencils oriented along the x direction, as in Fig. 1, 1D FFTs along this direction can be readily taken. For transforms in y, the subpencils need to be reoriented along the y direction, by taking a transpose locally within the plane on the CPU, before copying it over to the GPU. A highly efficient CUDA library call can be used to perform this transpose on the GPU and transfer data between GPU and CPU in a single operation. For transforms in z, communication

is required among all parallel processes to transpose x-y to x-z slabs of data. This involves packing and unpacking the data before and after the network communication. A more complex memory reordering is required while unpacking data received through network communication from other processes. We use optimized GPU *kernels* (called *zero-copy*) where GPU threads directly access data on the CPU without having to explicitly copy the data to the GPU [30]. These kernels are used only for unpacking, while the CUDA library call is used to pack the data, as *zero-copy* kernels use GPU resources for data transfers, which are required for fast computations.

In our "batched asynchronism" algorithm, as computed results for each subpencil in Fig. 1 are returned to the CPU, it is possible for communication calls, to be issued as soon as each pencil of data is ready. In principle this would allow some communication (among the CPUs) to occur simultaneously with data transfers and computations on the GPU. However, testing has shown that, consistent with the usual advantages of a small number of larger messages, communication of an entire slab scales better to larger problem sizes. Consequently we have designed the code such that, communication of computed results is initiated only when the computation is complete; and likewise the next round of computations is initiated only when a communication call providing data to be operated on has completed. This implies no overlap between operations involving the GPU and communication; but with GPU-CPU data transfer and computations overlapping each other in a highly efficient manner, this scheme still provides the best performance overall.

While our batched asynchronism algorithm as sketched out above is designed to handle the largest problem size that can be accommodated on the system, some remarks about factors that constraint the largest problem size feasible are appropriate. For example, we have actually timed the code up to $24\,576^3$ using 4096 nodes out of 4608 on Summit. However, the code eventually becomes less efficient, and smooth operation at that scale is less likely because of factors such as longer waiting times on a busy system, and greater susceptibility to hardware failures for jobs that require nearly the full system, in addition to reduced flexibility for on-the-fly processing that adds to memory requirements. Conversely, this also means, by aiming at a more realistic problem size, we are able to operate without much concern for memory limitations.

The maximal code performance that we obtained, as detailed in Ref. [16], can be briefly summarized as a speed up or GPU acceleration in the range of 3–5 relative to the best-performing CPU code on the same machine, and approximately 14.2 sec of wall clock per (second-order Runge-Kutta) time step using 3072 nodes of Summit, for simulation of the velocity field only, at $18\,432^3$ grid resolution. We use single-precision arithmetic in the performance runs and simulations in the following sections as the impact of machine precision on the results of the type we are interested in is weak [8]. Communication costs hold the key to further advancements on this or other yet more advanced platforms to come.

To conclude this computing-oriented section, we note that our focus has been on presenting both the overall principles and some specific design considerations necessary for achieving high performance at large problem sizes on one of the fastest CPU-GPU leadership-class supercomputers in the world at this time. Hardware and concepts defining leadership-class and extreme-scale computing can be expected to evolve rapidly in the future (as they have recently). However, it seems quite certain that optimization of data movement, and adaptability in the face of rapid changes in the high-performance computing environments, will continue to be crucial for future success.

III. MRIS: METHODOLOGY AND VALIDATION

In this section we begin with a more detailed discussion of the MRIS methodology and how this approach can be tested, for forced stationary isotropic turbulence. A validation study is presented which addresses single-point statistics, issues of statistical independence, and two-point statistics crucial to the material of Sec. IV. The forcing scheme we use is designed to reduce the statistical variability of spatially averaged statistics in time, by freezing the energy spectrum in the lowest few wave-number shells [31], at values derived from long-time averages of results from stochastic

forcing [32]. However, the MRIS methodology should be compatible with other forcing schemes that share the principle of maintaining energy by forcing at the large scales as well.

A. The MRIS approach and a validation procedure

As noted in Sec. I, as expectations for DNS rise due to a combination of scientific need and advances in computing power, a pressing challenge is that full-length simulations spanning several large-eddy timescales at extreme-scale resolution pushing the envelope of latest leadership-class platforms will likely be inaccessible. However, if a turbulent flow is statistically stationary and the focus is on small-scale phenomena with short timescales, we suggest that a much more viable alternative exists. In the proposed (MRIS) approach, we replace sampling over a long, continuous simulation at high resolution, say, N^3 by sampling over a number of short simulation segments that possess a demonstrable degree of statistical independence. Although a full-length N^3 resolution may be excessively costly, a long simulation at some lower resolution (say, N_1^3) can be assumed to be available, either from prior work or through new calculations. Our strategy is to use multiple snapshots, with approximate independence through a separation in time from an N_1^3 simulation, as initial conditions for the short N^3 segments, which are then disjoint from each other. Since the small scales adjust to grid refinement ($N_1^3 \to N^3$) rapidly, these segments need not be long. The total cost of simulations at N^3 will then be measured in Kolmogorov timescales instead of large-eddy timescales. This results in major cost savings, that will in turn make well-sampled results at high resolution much more readily feasible than otherwise.

It may be noted that while ensemble averaging over multiple independent simulations in turbulence is not common, it has been used before, in situations where statistical variability per simulation can be very substantial [33]. For us, a critical test for MRIS is whether the results are close, within some margins of uncertainty, to those from an actual, full-length N^3 simulation. For validation, we consider a full-length DNS that is available at some affordable value of N, at a Reynolds number sufficiently high to show clear intermittency, and is very well resolved in space and time—essentially, usable as a high-accuracy benchmark that MRIS results can be compared to. Resolution in space can be expressed by the nondimensional parameter $k_{\max}\eta$, where $k_{\max} = \sqrt{2}N/3$ is the highest wave number resolvable on an N^3 grid of length 2π units on each side, and $\Delta x/\eta \approx 2.96/k_{\max}\eta$. Accuracy in time may be controlled through the Courant number, which in the present flow without a mean velocity is defined as

$$C = \Delta t \left[\frac{|u|}{\Delta x} + \frac{|v|}{\Delta y} + \frac{|w|}{\Delta z} \right]_{\max}, \tag{3}$$

where u, v, w are velocity fluctuations, and the maximum is taken over all (N^3) grid points. In our second-order Runge-Kutta scheme a combination of $k_{\max}\eta \approx 1.4$ and $C = 0.6$ is usually adequate for low-order statistics but better resolution in both time and space are important for higher-order quantities strongly impacted by intermittency.

For a given well-resolved instantaneous snapshot at N^3 resolution, we can truncate down to N_1^3 by removing content at all Fourier modes with wave number higher than the value of $k_{\max}$ that corresponds to N_1^3 resolution. This removal of high-wave-number modes leads to an immediate decrease in various quantities, including $\langle\epsilon\rangle$, that contain substantial high-wave-number content. Next, we run an N^3 simulation segment with this truncated field as initial conditions, filling in the "extra" Fourier coefficients beyond the value of $k_{\max}$ of an N_1^3 grid with zeros. The desired outcome is for $\langle\epsilon\rangle$ to recover quickly to its original value in the (reference) N^3 simulation. For a given N, this "recovery time" is expected to increase with N/N_1, being longer (thus less economical) if N_1 is a very low resolution. In cases where prior data at resolutions $N/3$ or $N/4$ are conveniently available, it would be useful to reach the desired resolution via an intermediate stage such as $N_1 \to N_2$ followed by $N_2 \to N$. Incidentally the "recovery" process examined here has some parallels with the process by which the large scales can regenerate the small scales if the latter are artificially removed [34], provided the large scales are themselves maintained.

TABLE I. Selected parameters in simulation segments used for MRIS validation: from N_1^3 to N^3 (via N_2^3 if applicable), $k_{\max}\eta$ on N^3 grid, number of segments (M), time span in units of τ_η, and ensemble-averaged $\langle\epsilon\rangle$ and S_ϵ, at the beginning and end of each segment (subscripts b and e, respectively). Initial conditions for Cases 3 and 6 were taken from the end of Cases 2 and 5. In the reference 3072^3 simulation the time-averaged values of $\langle\epsilon\rangle$ and S_ϵ were 1.409 and 0.588, respectively. All simulations in this table were performed using a Courant number of $C = 0.25$, with the same forcing parameters and viscosity.

Case	N_1	N_2	N	$k_{\max}\eta$	M	β	$\langle\epsilon\rangle_b$	$S_{\epsilon b}$	$\langle\epsilon\rangle_e$	$S_{\epsilon e}$
1	768	–	3072	4.2	11	4	1.374	0.471	1.410	0.588
2	768	–	1536	2.1	22	2	1.375	0.471	1.409	0.585
3	–	1536	3072	4.2	22	2	1.409	0.585	1.410	0.587
4	384	–	1536	2.1	22	4	1.135	0.264	1.397	0.577
5	384	–	768	1.05	22	4	1.135	0.264	1.403	0.529
6	–	768	1536	2.1	22	2	1.403	0.529	1.404	0.586

Since we are investigating resolution effects, comparisons should be based mainly on quantities that are sensitive to the small scales. The list we consider includes the mean dissipation rate ($\langle\epsilon\rangle$), the dissipation skewness [35] (S_ϵ), and the energy spectrum at high wave number, as well as direct indicators of intermittency such as the statistics of dissipation rate and enstrophy fluctuations evaluated at a point or averaged locally in space.

For reference in the next three subsections, Table I shows important parameters for tests conducted in our MRIS validation study, with reference to a simulation at $R_\lambda \approx 390$ (one of the values tested in Ref. [8]), with $N = 3072$ at $k_{\max}\eta \approx 4.2$ which provides good resolution for the small scales. This full-length "reference" simulation was run for 5.5 T_E, with 22 snapshots written at intervals of 0.25 T_E apart. We truncate each snapshot down to 768^3 and examine how the numerical solutions recover if we (in Case 1) directly apply a 4 times increase in resolution back to 3072^3, or (Cases 2 and 3, combined) through two successive 2 times increases in resolution. We are interested in whether a new stationary state forms in a short period of time, with statistics closely resembling those extracted from the 3072^3 reference simulation. Similar tests (Cases 4 as well as 5 and 6) are also conducted to see if acceptable results can be obtained from poorly resolved velocity fields (in this case, 384^3 with $k_{\max}\eta$ as low as 0.5) in a similar manner. Although, through the definition of η, changes in $\langle\epsilon\rangle$ lead to changes in η and hence $k_{\max}\eta$ since $\eta \propto \langle\epsilon\rangle^{-1/4}$ this effect is weak even if long-time variations in the order of 10% [31] are considered. Simulations listed in this table were performed using CPU-based nodes on the 35-petaflops supercomputer Frontera at the Texas Advanced Computing Center.

B. Single-point statistics and spectra

Both Table I and Fig. 2 provide information on the dissipation rate and dissipation skewness, which can both be written explicitly in terms of the dissipation spectrum. When a substantial collection of high-wave-number modes is abruptly removed the dissipation rate drops, while subsequent transfer of energy from the large scales (which are forced) will allow a recovery. Since forcing is applied at the large scales we do not expect its details to affect the small-scale dynamics significantly [36]. The contrast between Cases 1 and 4 shows, as expected, that truncation at a lower wave-number cutoff leads to a stronger reduction of dissipation rate and a slower subsequent recovery. The route of two successive refinements (2 times each) requires fewer time steps to be run on the targeted finer (N^3) grid than a direct 4 times refinement—which translates to lower resource requirements overall. Similar but stronger trends are observed for the dissipation skewness, which contains more high-wave-number content than the mean dissipation rate.

In wave-number space, an immediate consequence of grid refinement is that energy can now be transferred to higher wave numbers that were not represented before. Figure 3 confirms that

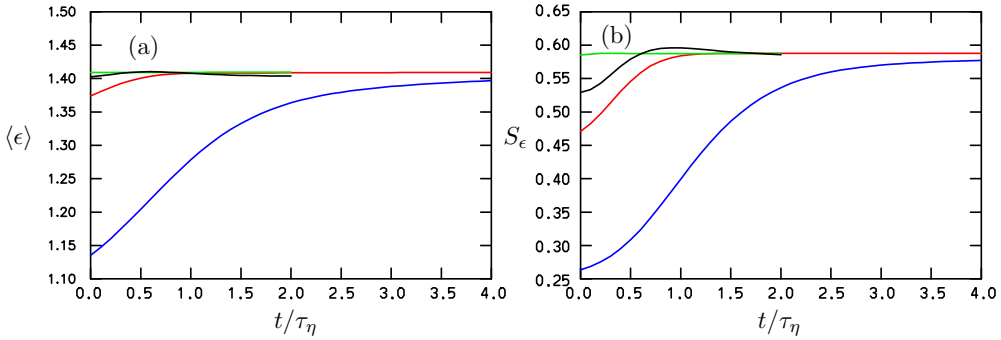

FIG. 2. Evolution of (a) $\langle\epsilon\rangle$ and (b) S_ϵ ensemble-averaged over multiple simulation segments, for Cases 1 (red), 3 (green), 4 (blue), and 6 (black), of different lengths as noted in Table I.

the small scales do adjust rapidly, with $E(k)$ at the end of the short simulation segments being nearly indistinguishable from results in the reference simulation. For Case 3, although the spectrum initially has a mild pileup at the $k_{\max}$ of the intermediate-sized (1536^3) grid (resulting from Case 2), a well-behaved functional form soon emerges.

Interest in the behavior of fluctuations of dissipation rate and enstrophy is a primary motivator for resolving the small scales as well as possible. In Fig. 4 we show information on the time history of [Fig. 4(a)] peak values (over all grid points) and [Fig. 4(b)] the probability density function (PDF) of normalized dissipation and enstrophy, obtained from the simulation segments of Case 3. In Fig. 4(a), despite substantial variability, the peak values can be seen to adjust to a new, stable stationary state, after only about 0.5 τ_η. The observed peak values in this new stationary state agree well with time-averaged values in the reference 3072^3 simulation (black dashed lines, partly hidden). Higher values of peak $\Omega/\langle\Omega\rangle$ also indicates enstrophy is more intermittent [8]. The dissipation PDF data at different times in Fig. 4(b) are also in support of a rapid approach to a new stationary state, consistent with the reference simulation.

C. Tests of statistical independence

The statistical quality of results from MRIS depends on the number of segments (M) available for ensemble averaging, and their degree of statistical independence. The latter is expected to be

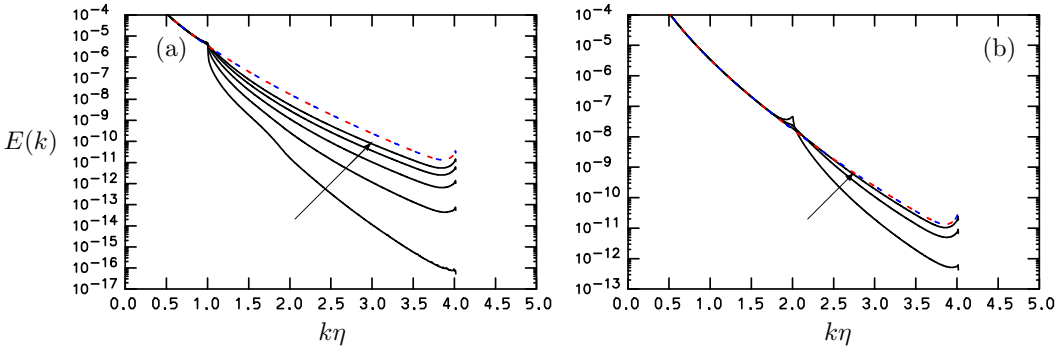

FIG. 3. Development of the energy spectrum as a result of grid refinement for (a) Case 1 and (b) Case 3 (per Table I). For clarity, only early-time data in the short segments (at increments of 0.1 τ_η, following the arrows) are shown. A blended red and blue dashed line gives spectra at the end of the short segments and time-averaged within the 3072^3 reference simulation.

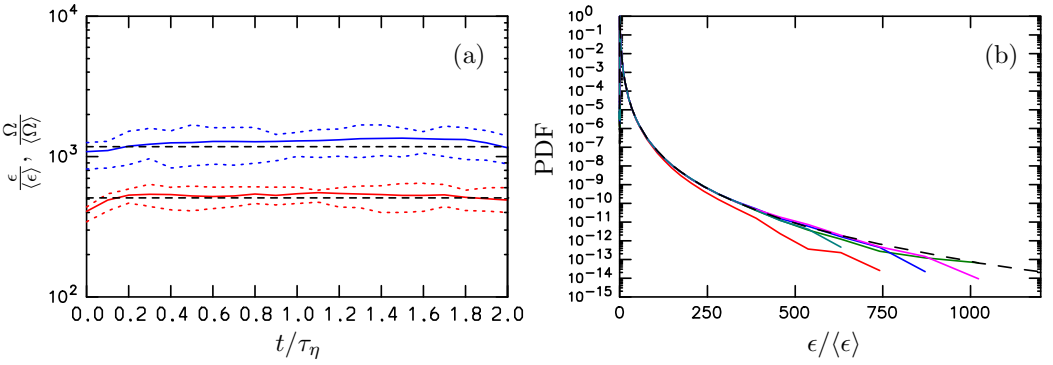

FIG. 4. Statistics of normalized dissipation and enstrophy obtained from multiple simulation segments for Case 3 in Table I. (a) Peak values: ensemble-averaged (solid lines) and 25th and 75th percentiles (dashed lines), red for dissipation, blue for enstrophy. (b) PDFs: red for data at $t = 0$, green, blue, magenta, cyan for $t/\tau_\eta = 0.5$, 1.0, 1.5, 2.0 respectively. In both frames black dashed lines (partly hidden) give results from the reference 3072^3 simulation for comparison.

a function of scale size, and closely related to the time separation (τ_0) between lower-resolution snapshots used as initial conditions for the MRIS segments, with the overall sampling period being effectively $\mathcal{T} = M\tau_0$. Statistical errors in DNS results can often be quantified via confidence intervals computed after the fact. However, it would be useful to develop some *a priori* estimates for the minimum τ_0 desired, depending on the nature of the quantity being sampled, and in relation to timescales τ_η or T_E. We explore this issue below using both one- and two-time statistics.

With stationary turbulence in mind, a basic question for one-time statistics, such as the volume-averaged energy dissipation rate ($\langle\epsilon\rangle$), is whether significant and random departures from the mean of either sign are consistently observed within a time period $\mathcal{T}$. If a signal shows persistent behaviors (such as monotonic variations) then the sampling period is too short; conversely, a predominance of rapid oscillations would suggest a small τ_0 is desirable, although strict independence over an interval of τ_0 is not necessary.

Figure 5 shows data from two long simulations at R_λ 390, of different resolutions as noted in the figure captions. The first is the one used to initiate high-resolution MRIS segments, whereas the

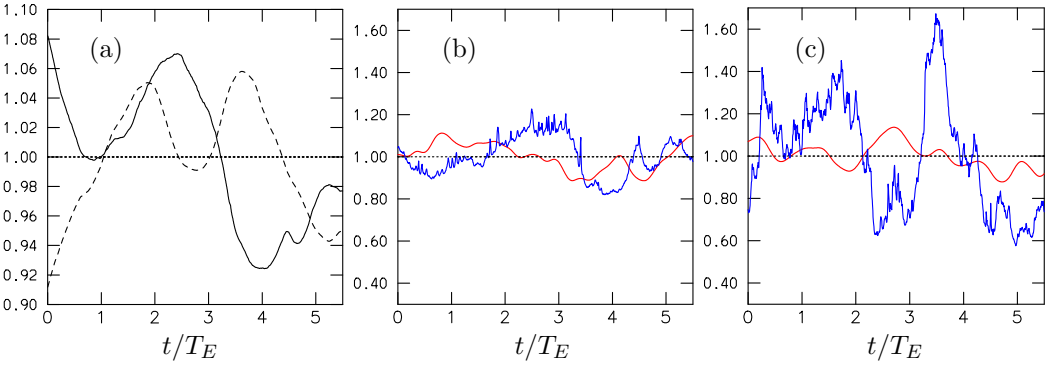

FIG. 5. (a) $\langle\epsilon\rangle/\langle\epsilon\rangle_T$ over a period of 5.5 T_E in R_λ 390 simulations at $k_{\max}\eta \approx 1.4$ (solid curve) and 4.2 (dashed curve) respectively, where the notation $\langle\cdot\rangle_T$ denotes a time average of volume-averaged quantities. (b) Energy spectrum $E(k)$ normalized by a time average, for $k = 6$ (red) and $k = 0.95\ k_{\max}$ (blue), from the $k_{\max}\eta \approx 1.4$ simulation. (c) Similar data, with $k_{\max}\eta$ 4.2.

second is the high-resolution reference simulation in our validation study. In Fig. 5(a) it can be seen that, in both simulations, dissipation varies to about the same degree (of order 10% or less), and with similar timescales. This behavior is not a surprise, since the mean dissipation rate is determined by the large scales, and the same forcing is used in both data sets. However, quantities at disparate scale sizes should behave differently. Figures 5(b) and 5(c) show that the energy spectrum $E(k)$ is indeed very dependent on wave number. At low wave number, the red lines in Figs. 5(b) and 5(c) show slow and modest variations. In contrast at high k (near $k_{\max}$ at each resolution) the lines in blue resemble rapid oscillations superimposed on a smooth background signal, which itself varies more strongly at high resolution. Incidentally, since $E(k)$ is (at high k) the sum of energies held in a large number of Fourier modes in a spectral shell, we can infer that individual Fourier modes vary in time even more rapidly than for the $E(k)$ values shown.

While one-time statistics show directly how different quantities evolve in time, it is tempting to ask if we can assess independence between two single-time snapshots, at times t_1 and $t_2 = t_1 + \tau$, by computing some statistical correlations. For example, one may consider the two-time correlator $\sigma(\tau) = \langle \epsilon(\mathbf{x}, t_1)\epsilon(\mathbf{x}, t_2)\rangle / \langle \epsilon^2 \rangle$ which is analogous to the two-point correlator in space related to intermittency exponents [37]. Another possible scale-dependent measure of statistical coupling may be the coherency spectrum defined by $\rho(k, \tau) = E_c(k, \tau)/\sqrt{E(k, t_1)E(k, t_2)}$ where $E_c(k, \tau)$ is the cospectrum between $\hat{\mathbf{u}}(\mathbf{k}, t_1)$ and $\hat{\mathbf{u}}(\mathbf{k}, t_2)$ in wave-number space. However, both of these quantities are subject to contamination by the "random-sweeping" effect [38], in which small-scale structures may be simply moved to another location as a result of advective transport by the large scales. Such an effect will cause an artificial drop of $\sigma(\tau)$ even if the turbulence were frozen. Likewise, since a coherency spectrum basically measures the phase coupling between Fourier-transformed quantities in wave-number space [39], random sweeping can also cause an artificial decrease of the coherency spectrum, especially at high Reynolds numbers.

Since random sweeping is an artifact of a fixed observer seeing differences in time while small-scale structures are swept along by the fluid, an alternative approach free of this effect is thus to consider the flow conditions experienced by an observer moving with the flow, i.e., to use a Lagrangian framework [40]. For a general flow variable q, we can define the Lagrangian two-time correlator as

$$\sigma_L(q; \tau) = \langle q^+(t) q^+(t + \tau)\rangle / \langle q^2 \rangle \tag{4}$$

where superscripts + denote Lagrangian quantities evaluated along the trajectories of fluid particles moving with the local fluid velocity. Clearly, $\sigma_L(q; \tau)$ is unity at $\tau = 0$ but approaches the ratio $\langle q \rangle^2 / \langle q^2 \rangle < 1$ when τ is large enough for $q^+(t)$ and $q^+(t + \tau)$ to be statistically independent. Subtracting $\langle q \rangle^2$ from both the numerator and denominator of $\sigma_L(q; \tau)$ gives the correlation coefficient $\rho_L(q; \tau)$, which approaches 0 at large τ for any q.

Figure 6 shows sample results in the two-time correlators [in Fig. 6(a)] and correlations [in Fig. 6(b)], for three choices of the quantity q being (1) u^2 (square of one fluctuation), (2) ϵ, and (3) its square, whose behavior mimics extreme events of very high amplitude. In this ordering, as the dominant scales are shifted towards quantities associated with increasingly smaller scales, it is not surprising that both measures of dependence or correlation decrease with time lag more rapidly. The discrepancy between the green lines for $\sigma_L(\epsilon; \tau)$ and its Eulerian counterpart $\sigma(\epsilon; \tau)$ confirms the importance of random sweeping, whose effect is strongest at small τ. The contrast between Eulerian and Lagrangian data here is also consistent with past comparisons between the statistics of Eulerian and Lagrangian time derivatives [41,42]. However at $\tau = 0.4\, T_E$ this discrepancy is mild, which is also expected, since the large-scale motions responsible for the sweeping are well-sustained only for a finite time interval. For the velocity, at $\tau / T_E = 0.4$, $\sigma_L(u^2; \tau)$ is not close to the asymptotic value of 1/3 (which assumes the velocity PDF to be Gaussian. A clearer view of the degree of independence that remains at this time lag is given by the Lagrangian correlation functions in Fig. 6(b), where a value of 0.1 for $\rho_L(\epsilon; \tau)$ suggests a high degree of independence from a practical viewpoint. Both panels show that ϵ^2 has short timescales, which become shorter yet as resolution is increased, consistent with the emergence of stronger extreme events of short

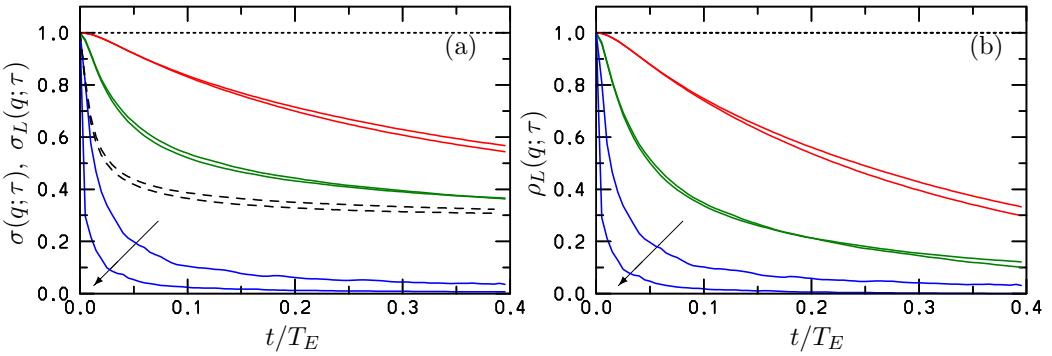

FIG. 6. (a) Eulerian (dashed lines) and Lagrangian (solid lines) two-time correlators versus time lag in units of T_E, for $q = u^2$ (red), ϵ (green), and ϵ^2 (blue). (b) Lagrangian two-time correlations, with same color coding for each variable. Both (a) and (b) show data obtained at two resolutions, $k_{\max}\eta \approx 1.4$ and 4.2, at R_λ 390. The only sensitivity evident is for ϵ^2 (resolution increasing in the direction of the arrows).

lifetimes. Finally, although data at only one Reynolds number is given in this figure, since the Lagrangian integral timescale of the dissipation rate decreases with respect to large-eddy timescales as the Reynolds number increases [43] it seems likely that $\tau_0/T_E \sim 0.4$ as a criterion for statistical independence of small-scale quantities will hold better yet at higher Reynolds numbers.

The assessment of resolution effects in Fig. 6 as discussed above suggests two high-resolution snapshots obtained by grid refinement from two modestly resolved ones will retain the degree of independence that originally existed in the former. Although this statement is understandably less valid for high amplitude events which are under-represented if the resolution is low, this supports the hypothesis that good sampling at high resolution can be derived from good sampling at modest resolution, in the MRIS approach that we propose.

D. Moments of 3D local averages

We now move to an examination of MRIS results for multipoint statistics in physical space—specifically, the scaling of moments of the 3D local averages of normalized dissipation rate and enstrophy, over scale sizes r ranging from the smallest (one grid spacing, Δx) to the largest (half of the length, L_0, of the periodic domain). Because our DNS is performed using Cartesian coordinates, we use 3D averaging over subcubes (instead of spheres). In the limit of $r \to 0$ the pth-order moment of $\epsilon_r/\langle\epsilon\rangle$ approaches $\langle\epsilon^p\rangle/\langle\epsilon\rangle^p$, which implies (for $p > 1$) small-scale resolution is crucial. In the other limit of $r \to \infty$ all moments approach unity, regardless of order, with homogeneity in space being the only requirement. However, the most important range of r is in the inertial range $\eta \ll r \ll L_1$, where the longitudinal integral length scale L_1 is about 0.2 L_0 in our simulations. In this range, classical refined similarity theory suggests

$$\left\langle \epsilon_r^p \right\rangle \Big/ \langle\epsilon\rangle^p \propto (r/\eta)^{-\zeta_p}, \tag{5}$$

where the dependence of the scaling exponents ζ_p (all positive) on the order p is of fundamental interest. Unfortunately since 3D averaging is challenging in both experiments and computation, many studies in the literature have, until recently [22], used instead 1D averages along a line, and/or a 1D surrogate [$(\partial u/\partial x)^2$, motivated by Taylor's frozen turbulence hypothesis], which is more intermittent than ϵ itself. Furthermore, accurate inferences of ζ_p require having a well-defined scaling range (hence a high Reynolds number) and attention to possible contamination from limitations in both resolution and sampling.

In our MRIS validation effort here, we focus on resolution and sampling. Figure 7 shows results averaged over multiple MRIS segments, for orders two to six (the latter being more demanding).

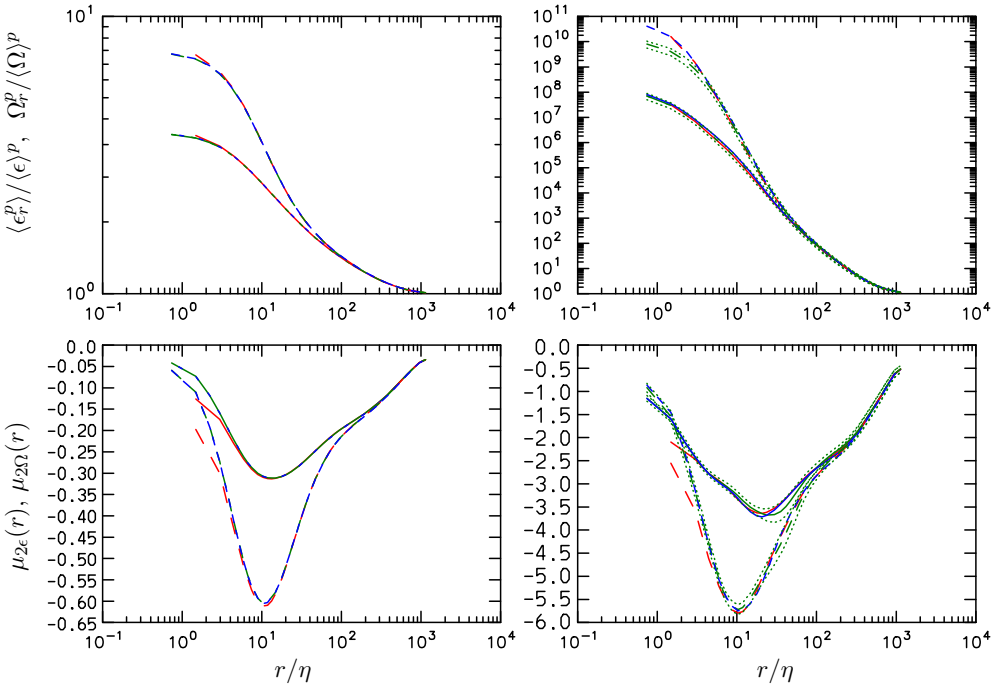

FIG. 7. Moments (top row) and logarithmic local slopes (bottom row) for 3D local averages of $\epsilon_r/\langle\epsilon\rangle$ (solid lines) and $\Omega_r/\langle\Omega\rangle$ (dashed lines), ensemble-averaged over multiple simulation segments from Cases 2 (red) and 3 (green). Lines in blue are from the reference 3072^3 simulation. Second moments on the left, sixth moments on the right. Dotted lines in green (very close to solid and dashed lines of the same color) show ±95% confidence interval for the sixth-order moments and local slopes.

Scaling exponents are estimated through logarithmic local slopes: i.e., $d\ln\langle\epsilon_r^p\rangle/d\ln r$, which would be equal to $-\zeta_p$ if a well-defined plateau exists. We introduce the notations $\mu_{p\epsilon}(r)$ and $\mu_{p\Omega}(r)$ for the local slopes for (the moments of) ϵ_r and Ω_r, respectively. Since the Reynolds number in our MRIS validation study is not high, it is not surprising that local slopes in this figure do not show a clear scaling range. Instead, there is a hint of an inflection point developing in the neighborhood of $r/\eta \sim O(100)$. Stronger intermittency in Ω_r versus ϵ_r is manifested clearly in higher values of the moments at small r, an effect that is noticeable up to $r/\eta \approx 200$. Values of the moments at small r increase very strongly with p, indicating that the resolution needed to observe flat plateaus as $r \to 0$ becomes harder to achieve. For the data shown, comparison between red and green lines suggests the effects of resolution are largely confined to $r/\eta \leqslant O(5)$, with very little apparent effect at intermediate scales close to the inflexion point noted above. This suggests $k_{\max}\eta \approx 2$ (as for the red lines) may be sufficient for investigating some aspects of inertial-range intermittency, although sufficient sampling is still necessary. With good sampling, very good agreement is seen between lines in green and blue: i.e., results on the local averages from the full-length 3072^3 reference simulation can be well recovered from much less-expensive data derived from MRIS (Case 3). We also observe that the data from the reference simulation, in blue, fall within the ±95% confidence interval lines for the sixth-order moments (and local slopes) of both dissipation and enstrophy. This shows good sampling from the MRIS approach is achieved and very good agreement with the reference simulation is observed within sampling uncertainties. The difference in sixth-order moments of locally averaged enstrophy at small r/η is likely due to the removal of Fourier modes contaminated by aliasing errors, as the velocity field from the reference simulation was initially truncated.

TABLE II. Parameters for production simulations at different Reynolds numbers, using the MRIS approach. All simulations in this table were performed using a Courant number of $C = 0.3$.

R_λ	N	$k_{\max}\eta$	β	M	$\langle\epsilon^2\rangle/\langle\epsilon\rangle^2$	$\langle\Omega^2\rangle/\langle\Omega\rangle^2$
390	1024	1.4	2	22	3.869	7.665
390	1536	2.1	2	22	4.034	7.938
390	3072	4.2	2	22	4.074	7.969
650	2048	1.4	2	15	4.357	8.718
650	3072	2.1	2	15	4.575	9.133
650	6144	4.2	2	15	4.664	9.214
1000	4096	1.4	2	10	4.949	9.901
1000	6144	2.1	2	10	5.250	10.556
1000	12 288	4.2	2	10	5.381	10.745
1300	12 288	3.0	1	10	6.103	12.238
1300	18 432	4.5	1	10	6.142	12.288

IV. MRIS: STUDY OF INTERMITTENCY AT HIGH RESOLUTION

A major motivation behind this paper has been a desire to contribute towards a high-fidelity characterization of both dissipation range and inertial range intermittency in high Reynolds number turbulence. This pursuit is very resource intensive, and large simulations that resolve the small scales well are necessary. The works described in Secs. II and III were in fact undertaken in order to identify a viable path towards meeting these challenges.

Table II shows the resolution levels and selected parameters of production simulations that have been performed using GPUs on Summit, combined with the MRIS approach starting from modest resolutions at four targeted Reynolds numbers. Results at R_λ 390 here are equivalent to those reported in the MRIS validation study of Sec. III. As resource requirements increase, the number of short simulation segments employed is fewer. Following estimates obtained in Sec. III, each segment is 2 τ_η long, except for those at highest Reynolds number on a $18\,432^3$ grid. In the latter case we decided to shorten each segment to 1 τ_η, partly because approach to a new stationary state in the manner of Fig. 4(a) took only about 1 τ_η, and partly because better overall sampling is likely from taking averages over more segments of shorter duration than over fewer segments of longer duration. Normalized second-order single-point moments in Table II are seen to increase systematically with both Reynolds number and resolution, while being higher for enstrophy than the dissipation. Sensitivity to resolution from $k_{\max}\eta \approx 2$ onwards appears to be relatively weak, thus suggesting, at least at $k_{\max}\eta \approx 4$, a certain degree of convergence has been reached.

The study of intermittency is a very broad subject, including the statistics of velocity gradients [44,45], velocity increments [22,46], use of multifractal theory [47,48], and various other aspects. For reasons of the scope and length of this paper we will just focus here on the moments of the locally averaged dissipation rate (ϵ_r) and enstrophy (Ω_r), and their statistical relationships to each other. The moments of ϵ_r directly enter into a number of intermittency corrections based on the Refined Similarity Hypothesis [11,21,49]. For instance, the Refined Similarity prediction for the nth-order velocity structure function is of the form

$$D_n(r) = C_n\langle\epsilon_r^{n/3}\rangle r^{n/3}, \tag{6}$$

where C_n are universal constants. The moments of Ω_r provide a useful contrast, as well as information on the structural differences between strain-dominated and rotation-dominated regions in the instantaneous turbulent flow.

Figure 8 shows data on second and fourth moments from the highest resolution simulations (all with $k_{\max}\eta \geqslant 4$) available at all four Reynolds numbers, in a manner similar to that of Fig. 7. In

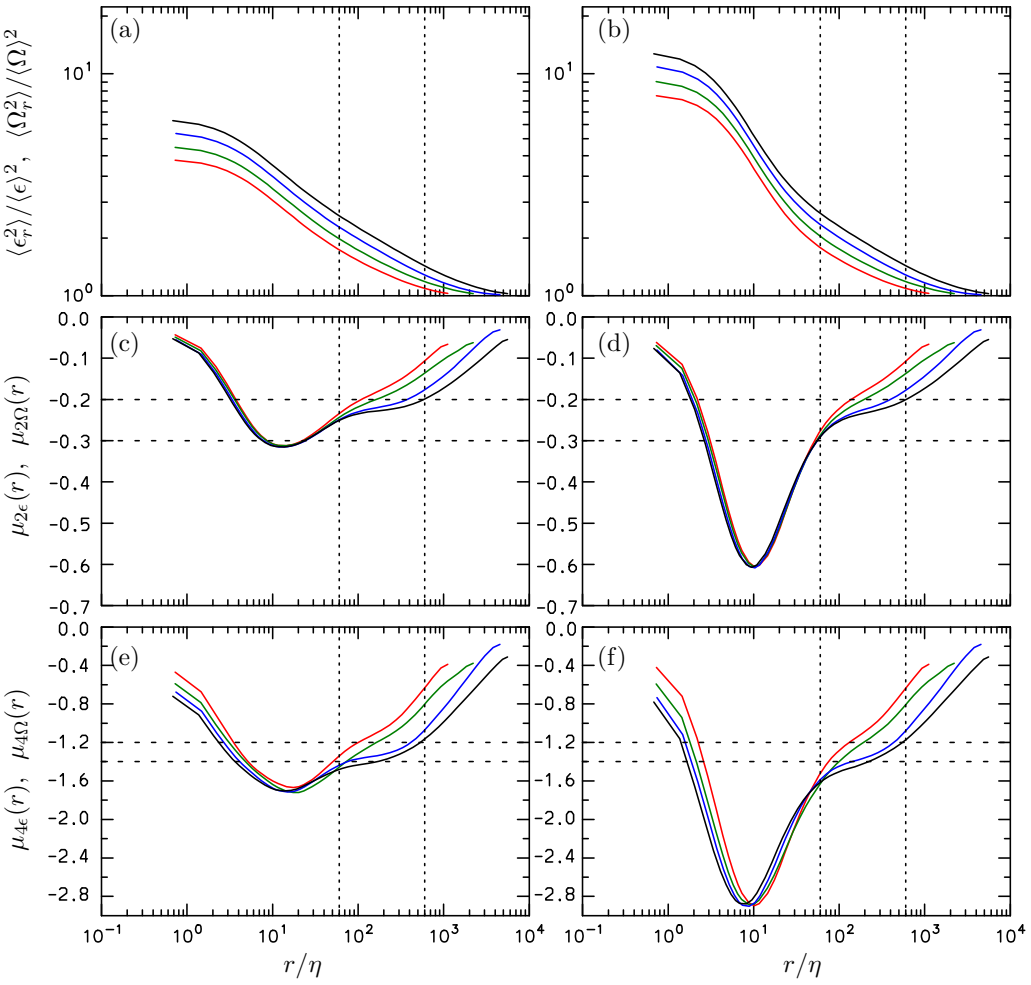

FIG. 8. Ensemble-averaged normalized second-order moments (a), (b) of 3D local averages of dissipation rate, ϵ_r (left), and enstrophy, Ω_r (right), from simulations at highest resolution available at each R_λ. Ensemble average of the logarithmic local slopes of the second-order (c), (d) and fourth-order (e), (f) moments of local averages. The different colors correspond to different R_λ: 390 (red), 650 (green), 1000 (blue), and 1300 (black). Horizontal dashed lines in panels (c)–(f) are included to assist in inference of scaling exponents from the graphs, at the highest R_λ.

principle, local slopes should smoothly approach zero at both the small r and large r limits, scaling with η for the former but L_1 for the latter. For the second moment, this scaling at small r explains why the local scopes are nearly independent of Reynolds number up to r/η at least about 10, while the scaling at large r explains why, with $L_1/\eta \propto {R_\lambda}^{3/2}$ according to classical scaling, the local slope curves eventually diverge at intermediate scale ranges in the manner shown.

In Fig. 8 we have included two vertical dotted lines, at r/η 60 and 600, which have been proposed [22] as approximate bounds for inertial range scaling where applicable. It can be seen that as Reynolds number increases, an inflexion point gradually develops into a plateau, which is somewhat flatter for dissipation than enstrophy. The values of the exponents $\mu_{2\epsilon}$ and $\mu_{2\Omega}$ appear to differ only very slightly, with both being close to 0.23. This difference appears to be less than what past experimental data based on 1D surrogates averaged along a line suggested [48,50,51]. On the other

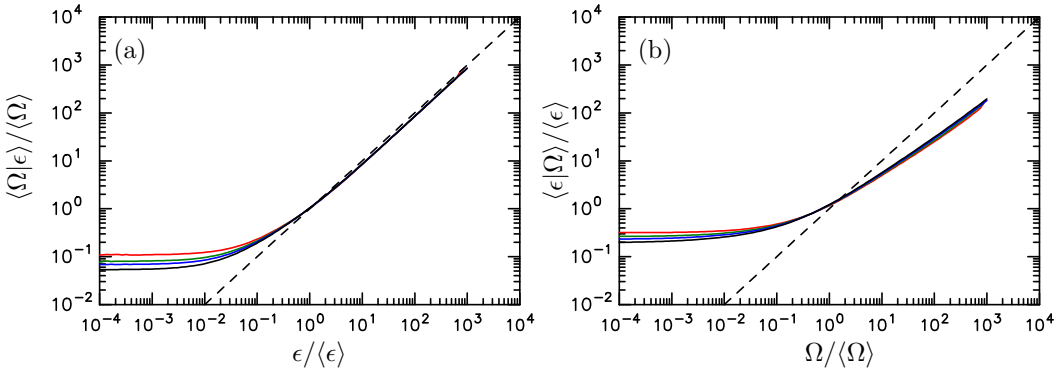

FIG. 9. First-order conditional moments of (a) enstrophy given dissipation rate and (b) dissipation given enstrophy at R_λ 390 (red), 650 (green), 1000 (blue), and 1300 (black). Dashed line of slope 1 corresponds to enstrophy and dissipation scaling similarly.

hand, greater intermittency in the dissipation range for enstrophy compared to dissipation implies local slopes at smaller r/η are of larger magnitude than those for dissipation (most significantly at $r/\eta \approx 10$ in the figure), while homogeneity ultimately forces both sets of curves to agree with each other at sufficiently large r. Further investigations are appropriate in the future, especially when data at yet higher Reynolds numbers with a comparable degree of resolution become available.

Curves for local slopes for the fourth-order moments shown in the bottom row of the figures are of generally similar shape when compared with those for the second-order moments. However, as can be expected, differences at small r indicate the small-scale resolution in this case is less satisfactory, especially at higher Reynolds numbers. Careful observation in the nominal inertial range of r/η also indicates inertial range behavior is less clearly developed at fourth order, while the difference between dissipation and enstrophy in the same range is more significant than that seen for the second moment.

A recurrent question in the study of intermittency is whether the dissipation rate and enstrophy, as quadratic invariants of the symmetric and antisymmetric parts, respectively, of the velocity gradient tensor, possess the same scaling properties [52] or even scale together [53]. For an update on this question we present some conditional moments derived from the present database. We note that conditional statistics given dissipation and or enstrophy have been used recently to study vortex stretching [54]. Figure 9 shows the (single-point) conditional mean of [Fig. 9(a)] enstrophy given the dissipation, and [Fig. 9(b)] dissipation given the enstrophy, at four Reynolds numbers. While samples where the conditioning variable up to nearly 10^4 in magnitude do exist, we show results only up to 10^3 on the x axes since data beyond that are noisy. The present results are similar to those in a previous investigation [55] at low to moderate values of the conditioning variable, but more accurate at a high conditioning value of the enstrophy. The data indicate that a high ϵ is likely to be accompanied by a high Ω; but in contrast a high Ω is likely to be accompanied by a ϵ which, although still large, may be nearly an order of magnitude smaller. At the other extreme of very low dissipation or enstrophy both of the conditional means are relatively flat, with a weak trend of decrease with increasing Reynolds number. This suggests, in the limit of vanishingly small dissipation or enstrophy, both variables become independent of each other while being mostly substantially below their average intensities. This observation is also consistent with results on joint probability density functions presented in Ref. [53].

In Fig. 10 we extend results on conditional means to moments of different orders and to local averages over volumes of linear size associated with the dissipation and inertial ranges. To facilitate the comparisons, for each $p > 1$ we have taken the pth root of the moment. For a given ϵ_r, and as

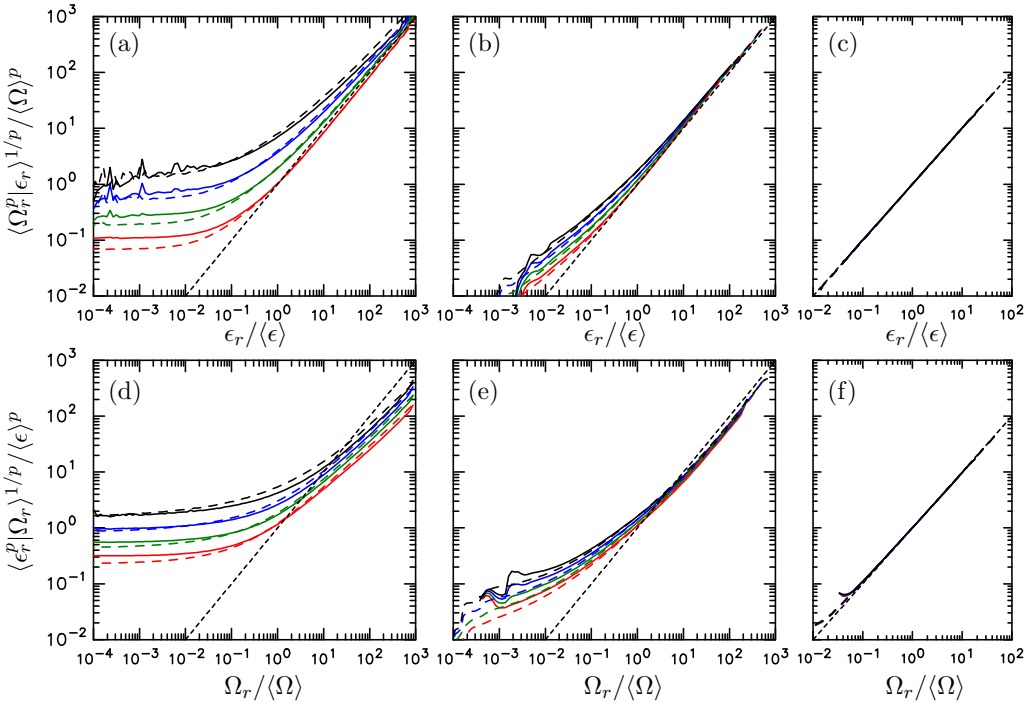

FIG. 10. Conditional moments of order p of local averages of enstrophy given local averages of dissipation rate (top) and vice versa (bottom) for (a), (d) $r/\eta \approx 0.7$, (b), (e) $r/\eta \approx 11$, and (c), (f) $r/\eta \approx 90$. First, second, third- and fourth-order moments are shown by curves in red, green, blue, and black. Solid lines from simulations at $R_\lambda \approx 390$ and dashed lines from $R_\lambda \approx 1000$. Dashed line of slope 1 corresponds to enstrophy and dissipation scaling together.

order p increases, moments of the conditional samples of Ω_r become increasingly dominated by samples that turn out to be very large. This explains, for instance, in the top half of the figure, why the black lines lie uniformly higher than the blue, as can be seen in Figs. 10(a) and 10(b). Effects of Reynolds number on these joint statistics appear to be weak. On the other hand, in Fig. 10(c), data for moments of all orders all collapse upon the line that indicates $\langle \Omega_r | \epsilon_r \rangle = \epsilon_r$. In the bottom half of this figure we show conditional moments of dissipation given the enstrophy. It can be seen that, as r/η approaches the inertial range, both sets of conditional moments, for all Reynolds numbers, and at all orders, largely collapse together on the diagonal line that would be satisfied also if the two locally averaged variances were to take the same values. This behavior suggests that ϵ_r and Ω_r do, to a good approximation, scale together in the inertial range.

It should be apparent that the results reported in this section, involving simulations at $12\,288^3$ and $18\,432^3$ resolution, have required use of substantial computational power which is itself in high demand. Recalling considerations in Sec. III, in the case of R_λ 1300, with the ratio $T_E/\tau_\eta \approx 136$ (based on Ref. [18]) a simulation of 5.5 T_E in length similar to the MRIS validation study earlier in this paper will be 748 τ_η in length. In contrast, if we were to obtain 22 simulation segments (same number as in Table II) only 1 τ_η each in length (based on Table III) then the cost would be roughly equivalent to 22 τ_η. This is a factor of 34 reduction in resource requirements—changing hypothetical periods of nonstop computing from months (which, incidentally, is not allowed) to days, thus making a great impact on the feasibility of the computations.

V. CONCLUSIONS AND DISCUSSION

In this paper we have reported on advances in developing, and actually applying, a capability of performing direct numerical simulations (DNS) of turbulence at extreme-scale problem sizes, that would otherwise be impossible or impractical in their resource requirements. The challenges faced here have arisen due to the fact that, despite dramatic advances in world-class computational resources, insatiable demands for high Reynolds number, improved small-scale resolution, and other needs, are, ironically, pointing to increasing challenges for researchers' abilities to conduct long simulations at leadership-class problem sizes.

The first innovation reported is our recent success in development of a parallel algorithm for DNS of incompressible turbulence on a 3D periodic domain, which successfully scaled up to a grid resolution of $18\,432^3$ (more than 6 trillion) grid points on a CPU-GPU machine which is currently one of the fastest supercomputers in the world. Key features of this algorithm involve taking full advantage of memory on both the CPU and GPU, in a manner that presents new opportunities for asynchronous execution involving overlapping of computation on multiple GPUs and data movement between the CPU and the GPUs. While this algorithm has some machine-dependent elements, this effort may be viewed as a case study for future exascale computing, where major efforts in adapting or even rewriting codes for a top-ranked machine will likely be required.

Despite the algorithmic advancement noted above, we point out that since resource requirements simulating an N^3 problem for a prescribed period of time increases at least as fast as N^4, full-length simulations spanning multiple large-eddy timescales at "leadership-class" problem sizes are essentially impossible. However, if the prime interest is in small-scale motions of short timescales, we show that a much more viable alternative exists, in an approach here termed Multiple Resolution Independent Simulations (MRIS). The essence of MRIS is to first perform a (much less costly) simulation at low or modest resolution, take multiple snapshots well separated in time, and refine the grid to obtain multiple short simulation segments at the highest resolution. With appropriate attention given to statistical independence, we have shown that ensemble averaging over a number of such short segments produces results essentially equivalent to sampling from a long simulation with samples separated from each other by fractions of a large-eddy timescale. In this paradigm the total cost of a simulation of stationary isotropic turbulence at very high resolution can be measured in (multiples of) Kolmogorov timescales rather than eddy-turnover times, resulting in tremendous savings at high Reynolds numbers. A validation study involving several single- and multipoint diagnostics as presented in this paper has apparently been successful. In particular, results in Figs. 3, 4, and 7 provided several examples of small-scale statistics obtained from the MRIS procedure being a close match with those taken directly for a full-length reference simulation at high resolution.

We have been able to apply the MRIS approach to obtain high-fidelity results concerning intermittency in both dissipative and inertial scale ranges in isotropic turbulence at four Reynolds numbers ranging from 390 to 1300 based on the Taylor scale. This work has provided an opportunity to overcome some of the limitations due to resolution and sampling in previous efforts. In particular, although reliable statistics on higher order moments had been difficult to achieve, results shown in Sec. IV are very robust, showing the benefits of leadership-class computing power applied productively. Calculations based on the statistics of 3D local averages of the dissipation rate and enstrophy show that although these quantities scale differently throughout the dissipation range, their inertial range properties are much more (although not exactly) similar. Conditional statistics also suggest strongly that these two variables do, to a good approximation, scale together when in the inertial range. The high fidelity of results obtained in this paper gives rise readily to the search for further physical insights, such as how differences and similarities in locally averaged dissipation and enstrophy may be connected to the incidence of canonical flow structures such as local shear layers [56] and vortex filaments of finite size.

In summary, we believe the body of work described in this paper provides a useful perspective concerning how turbulence researchers may be able to truly use emerging exascale platforms to the

fullest, and the challenges that the community can expect to face as well. For instance, machine architecture can influence choice of problem sizes (in our case, a factor of 6 in the number of grid points in each direction), as well as provide unique opportunities for asynchronism that requires serious rethinking of basic algorithmic principles. Optimizing communication and data transfer on heterogeneous machines will continue to be major challenges. The MRIS approach in this work has been developed to address the issue of how large simulations of limited time span imposed by practical constraints on resource availability can be designed to meet specific scientific needs. Other than physical problems where early-time phenomena are of greatest interest, the MRIS approach is likely to be applicable to studies of dissipation rate and fine-scale structure in passive scalar fields [31,57], as well as the fluid particle acceleration [58], which are both dominated by intermittency and characterized by short timescales.

ACKNOWLEDGMENTS

The authors are grateful to Prof. K. R. Sreenivasan and Prof. S. B. Pope for many valuable discussions. We also thank two anonymous referees, as well as Prof. T. Gotoh, Prof. Y. Kaneda, and Prof. A. Pumir for their valuable input on an earlier version of this paper. This research used primarily supercomputing resources granted through the 2019 INCITE and Summit Early Science programs, at the Oak Ridge Leadership Computing Facility, which is a U.S Department of Energy Office of Science User Facility supported under Contract DE-AC05-00OR22725. Computational resources were also provided by the Texas Advanced Computing Center, under the auspices of the National Science Foundation (NSF)'s Leadership Resource Allocations program, where we received support as a supplement to Grant No. 1510749 from the Fluid Dynamics Program at NSF.

[1] P. Moin and K. Mahesh, Direct numerical simulation: A tool in turbulence research, Annu. Rev. Fluid Mech. **30**, 539 (1998).

[2] T. Ishihara, T. Gotoh, and Y. Kaneda, Study of high-Reynolds number isotropic turbulence by direct numerical simulation, Annu. Rev. Fluid Mech. **41**, 165 (2009).

[3] T. Ishihara, K. Morishita, M. Yokokawa, A. Uno, and Y. Kaneda, Energy spectrum in high-resolution direct numerical simulation of turbulence, Phys. Rev. Fluids **1**, 082403 (2016).

[4] M. P. Clay, D. Buaria, T. Gotoh, and P. K. Yeung, A dual communicator dual grid-resolution algorithm for petascale simulations of turbulent mixing at high Schmidt number, Comput. Phys. Commun. **219**, 313 (2017).

[5] A. Gruber, E. S. Richardson, K. Aditya, and J. H. Chen, Direct numerical simulations of premixed and stratified flame propagation in turbulent channel flow, Phys. Rev. Fluids **3**, 110507 (2018).

[6] S. Jagannathan and D. A. Donzis, Reynolds and Mach number scaling in solenoidally-forced compressible turbulence using high-resolution direct numerical simulations, J. Fluid Mech. **789**, 669 (2016).

[7] J. Schumacher, J. D. Scheel, D. Krasnov, D. A. Donzis, V. Yakhot, and K. R. Sreenivasan, Small-scale universality in fluid turbulence, Proc. Natl. Acad. Sci. USA **111**, 10961 (2014).

[8] P. K. Yeung, K. R. Sreenivasan, and S. B. Pope, Effects of finite spatial and temporal resolution on extreme events in direct numerical simulations of incompressible isotropic turbulence, Phys. Rev. Fluids **3**, 064603 (2018).

[9] M. Lee and R. D. Moser, Direct numerical simulation of turbulent channel flow up to $Re_\tau \approx 5200$, J. Fluid Mech. **774**, 395 (2015).

[10] T. Watanabe, J. J. Riley, S. M. de Bruyn Kops, P. J. Diamessis, and Q. Zhou, Turbulent/non-turbulent interfaces in wakes in stably stratified fluids, J. Fluid Mech. **797**, R1 (2016).

[11] U. Frisch, *Turbulence: The Legacy of A. N. Kolmogorov* (Cambridge University Press, Cambridge, 1995).

[12] K. R. Sreenivasan and R. A. Antonia, The phenomenonology of small-scale turbulence, Annu. Rev. Fluid Mech. **29**, 435 (1997).

[13] M. P. Clay, D. Buaria, P. K. Yeung, and T. Gotoh, GPU acceleration of a petascale application for turbulent mixing at high Schmidt number using OpenMP 4.5, Comput. Phys. Commun. **228**, 100 (2018).
[14] D. Rosenberg, P. D. Mininni, R. Reddy, and A. Pouquet, GPU parallelization of a hybrid pseudospectral geophysical turbulence framework using CUDA, Atmosphere **11**, 178 (2020).
[15] M. K. Verma, R. Samuel, S. Chatterjee, S. Bhattacharya, and A. Asad, Challenges in fluid flow simulations using exascale computing, SN Comput. Sci. **1**, 178 (2020).
[16] K. Ravikumar, D. Appelhans, and P. K. Yeung, GPU acceleration of extreme scale pseudospectral simulations of turbulence using asynchronism, in *Proceedings of the International Conference for High Performance Computing, Networking and Storage Analysis (SC'19)* (ACM Press, New York, 2019), pp. 1–22, https://doi.org/10.1145/3295500.3356209.
[17] Y. Kaneda, T. Ishihara, M. Yokokawa, T. Itakura, and A. Uno, Energy dissipation rate and energy spectrum in high resolution direct numerical simulations of turbulence in a periodic box, Phys. Fluids **15**, L21 (2003).
[18] P. K. Yeung, X. M. Zhai, and K. R. Sreenivasan, Extreme events in computational turbulence, Proc. Natl. Acad. Sci. USA **112**, 12633 (2015).
[19] V. Yakhot and K. R. Sreenivasan, Anomalous scaling of structure functions and dynamic constraints on turbulence simulations, J. Stat. Phys. **121**, 823 (2005).
[20] J. Schumacher, K. R. Sreenivasan, and P. K. Yeung, Very fine structures in scalar mixing, J. Fluid Mech. **531**, 113 (2005).
[21] A. N. Kolmogorov, A refinement of previous hypotheses concerning the local structure of a viscous incompressible fluid, J. Fluid Mech. **13**, 82 (1962).
[22] K. P. Iyer, K. R. Sreenivasan, and P. K. Yeung, Refined similarity hypotheses using 3D local averages, Phys. Rev. E **92**, 063024 (2015).
[23] M. Lee, R. Ulerich, N. Malaya, and R. D. Moser, Experiences from leadership computing in simulations of turbulent fluid flows, Comput. Sci. Eng. **16**, 24 (2014).
[24] S. A. Orszag, Numerical methods for the simulation of turbulence, Phys. Fluids **12**, II–250 (1969).
[25] C. Canuto, M. Y. Hussaini, A. Quateroni, and T. A. Zang, *Spectral Methods in Fluid Dynamics* (Springer-Verlag, New York, 1987).
[26] P. D. Mininni, D. Rosenberg, R. Reddy, and A. Pouquet, A hybrid MPI-OpenMP scheme for scalable parallel pseudospectral computations for fluid turbulence, Parallel Comput. **37**, 316 (2011).
[27] G. S. Patterson and S. A. Orszag, Spectral calculations of isotropic turbulence: Efficient removal of aliasing interactions, Phys. Fluids **14**, 2538 (1971).
[28] R. S. Rogallo, Numerical experiments in homogeneous turbulence, NASA Tech. Memo 81315, NASA Ames Research Center, Moffett Field, CA (1981).
[29] D. Pekurovsky, P3DFFT: A framework for parallel computations of Fourier transforms in three dimensions, SIAM J. Sci. Comput. **34**, C192 (2012).
[30] IBM POWER9 NPU team, Functionality and performance of NVLink with IBM POWER9 processors, IBM J. Res. Devel. **62**, 9:1 (2018).
[31] D. A. Donzis and P. K. Yeung, Resolution effects and scaling in numerical simulations of passive scalar mixing in turbulence, Physica D **239**, 1278 (2010).
[32] V. Eswaran and S. B. Pope, An examination of forcing in direct numerical simulations of turbulence. Comput. Fluids **16**, 257 (1988).
[33] M. R. Overholt and S. B. Pope, Direct numerical simulation of a passive scalar with imposed mean gradient in isotropic turbulence, Phys. Fluids **8**, 3128 (1996).
[34] K. Yoshida, J. Yamaguchi, and Y. Kaneda, Regeneration of Small Scales by Data Assimilation in Turbulence, Phys. Rev. Lett. **94**, 014501 (2005).
[35] R. M. Kerr, Higher-order derivative correlations and the alignment of small-scale structures in isotropic numerical turbulence, J. Fluid Mech. **153**, 31 (1985).
[36] K. R. Sreenivasan, An update on the energy dissipation rate in isotropic turbulence, Phys. Fluids **10**, 528 (1998).
[37] K. R. Sreenivasan, An update on the intermittency exponent in turbulence, Phys. Fluids A **5**, 512 (1993).

[38] H. Tennekes, Eulerian and Lagrangian time microscales in isotropic turbulence, J. Fluid Mech. **67**, 561 (1975).
[39] P. K. Yeung, Multi-scalar triadic interactions in differential diffusion with and without mean scalar gradients, J. Fluid Mech. **321**, 235 (1996).
[40] P. K. Yeung, Lagrangian investigations of turbulence, Annu. Rev. Fluid Mech. **34**, 115 (2002).
[41] P. K. Yeung and S. B. Pope, Lagrangian statistics from direct numerical simulations of isotropic turbulence, J. Fluid Mech. **207**, 531 (1989).
[42] A. Tsinober, P. Vedula, and P. K. Yeung, Random Taylor hypothesis and the behavior of local and convective accelerations in isotropic turbulence, Phys. Fluids **13**, 1974 (2001).
[43] P. K. Yeung, S. B. Pope, and B. L. Sawford, Reynolds number dependence of Lagrangian statistics in large numerical simulations of isotropic turbulence, J. Turbul. **7**, N58 (2006).
[44] D. Buaria, A. Pumir, E. Bodenschatz, and P. K. Yeung, Extreme velocity gradients in turbulent flows, New J. Phys. **21**, 043004 (2019).
[45] R. Das and S. S. Girimaji, On the Reynolds number dependence of velocity-gradient structure and dynamics, J. Fluid Mech. **861**, 163 (2019).
[46] K. P. Iyer, K. R. Sreenivasan, and P. K. Yeung, Reynolds number scaling of velocity increments in isotropic turbulence, Phys. Rev. E **95**, 021101 (2017).
[47] C. Meneveau, K. R. Sreenivasan, P. Kailasnath, and M. S. Fan, Joint multifractal measures: Theory and applications to turbulence, Phys. Rev. A **41**, 894 (1990).
[48] C. Meneveau and K. R. Sreenivasan, The multifractal nature of turbulent energy-dissipation, J. Fluid Mech. **224**, 429 (1991).
[49] A. M. Obukhov, Some specific features of atmospheric turbulence, J. Fluid Mech. **13**, 77 (1962).
[50] F. Anselmet, R. A. Antonia, and L. Danaila, Turbulent flows and intermittency in laboratory experiments, Planet. Space Sci. **49**, 1177 (2001).
[51] S. Chen, K. R. Sreenivasan, and M. Nelkin, Inertial Range Scalings of Dissipation and Enstrophy in Isotropic Turbulence, Phys. Rev. Lett. **79**, 1253 (1997).
[52] M. Nelkin, Enstrophy and dissipation must have the same scaling exponents in the high Reynolds number limit of fluid turbulence, Phys. Fluids **11**, 2202 (1999).
[53] P. K. Yeung, D. A. Donzis, and K. R. Sreenivasan, Dissipation, enstrophy and pressure statistics in turbulence simulations at high Reynolds numbers, J. Fluid Mech. **700**, 5 (2012).
[54] D. Buaria, E. Bodenschatz, and A. Pumir, Vortex stretching and enstrophy production in high Reynolds number turbulence, Phys. Rev. Fluids **5**, 104602 (2020).
[55] D. A. Donzis, P. K. Yeung, and K. R. Sreenivasan, Energy dissipation rate and enstrophy in isotropic turbulence: Resolution effects and scaling in direct numerical simulations, Phys. Fluids **20**, 045108 (2008).
[56] T. Ishihara, Y. Kaneda, and J. C. R. Hunt, Thin shear layers in high Reynolds number turbulence—DNS results, Flow, Turbul. Combust. **91**, 895 (2013).
[57] K. P. Iyer, J. Schumacher, K. R. Sreenivasan, and P. K. Yeung, Steep Cliffs and Saturated Exponents in Three-Dimensional Scalar Turbulence, Phys. Rev. Lett. **121**, 264501 (2018).
[58] P. K. Yeung, S. B. Pope, E. A. Kurth, and A. G. Lamorgese, Lagrangian conditional statistics, acceleration and local relative motion in numerically simulated isotropic turbulence, J. Fluid Mech. **582**, 399 (2007).