

# Sensitivity of numerical simulations of turbulence to lower floating-point precision

By M. Karp<sup>†</sup>, R. Stanly<sup>†</sup>, H. Song, T. Mukha<sup>‡</sup>, L. Galimberti<sup>‡¶</sup>, S. Toosi<sup>||</sup>,  
L. Dalcin<sup>‡</sup>, S. Rezaeiravesh<sup>††</sup>, M. Münsch<sup>||</sup>, N. Jansson<sup>†</sup>, S. Markidis<sup>†</sup>,  
M. Parsani<sup>‡</sup>, S. T. Bose<sup>‡‡</sup>, S. K. Lele AND P. Schlatter<sup>†||</sup>

Modern computing clusters offer specialized hardware with reduced-precision arithmetic that can speed up the time to solution significantly. This is possible due to a decrease in data movement, as well as the ability to perform arithmetic operations at a faster rate. However, for high-fidelity simulations of turbulence, the impact of reduced precision on the computed solution and the uncertainty it introduces across flow solvers and different flow cases have not been explored in detail. This limits the optimal utilization of new high-performance computing systems, which often feature hardware support for lower floating-point precision. In this work, the effect of reduced precision is studied using four computational fluid dynamics (CFD) solvers (two incompressible, Neko and Simson, and two compressible, PadeLibs and SSDC) using three test cases: turbulent channel flow at  $Re_\tau = 550$  and higher, forced transition in a channel, and flow over a cylinder at  $Re_D = 3900$ . We observe that the flow physics is remarkably robust under the influence of lower floating-point precision, and that it behaves differently than spatial convergence errors. Our results indicate that different terms in the Navier–Stokes equations can be computed to a lower floating-point accuracy, and that single precision can be used effectively for the entirety of the simulation, showing no significant discrepancies from double-precision results across the solvers and cases considered. The exact point at which the lower numerical precision becomes dominant in comparison to other sources of uncertainty for turbulent flows is difficult to determine. This motivates further analysis and more nuanced tests of the quality of numerical simulations of turbulence.

---

## 1. Introduction

This study assesses the sensitivity of simulations of turbulent flows to reductions in numerical precision. While the sensitivity and impact of numerical discretization in space and time have been considered previously (e.g. Rezaeiravesh *et al.* 2021), the impact of floating-point representation across flow cases and different flow solvers has not been studied extensively in a systematic and objective way. To this end, we assess three different canonical flow cases often considered with high-fidelity simulations, namely transition, turbulence, and separation, across four different numerical solvers. This approach gives the opportunity to find the commonalities among the different discretizations and implementations, and to evaluate what terms in the Navier–Stokes equations are most sensitive

<sup>†</sup> KTH Royal Institute of Technology, Sweden

<sup>‡</sup> King Abdullah University of Science and Technology, Saudi Arabia

<sup>¶</sup> Politecnico di Milano, Italy

<sup>||</sup> Friedrich–Alexander–Universität (FAU) Erlangen–Nürnberg, Germany

<sup>††</sup> University of Manchester, United Kingdom

<sup>‡‡</sup> Cadence Design Systems, Inc.

to a reduction in numerical precision. Lower floating-point units are becoming commonplace in recent computer architectures and enable significant performance improvements and cost savings. As such, a thorough study of the applicability of lower floating-point precision for turbulence simulations is timely and gives guidelines on what terms in the Navier–Stokes equations may be efficiently computed in lower precision.

## 2. Background and methodology

The issue of floating-point formats in numerical simulation is not new but has conventionally been studied less than modeling, spatial, and temporal discretization errors (Homann *et al.* 2007). In general, for integrating the Navier–Stokes equations in time, the default has been to opt for the higher floating-point format available in hardware, namely the double-precision format defined by the IEEE standard. However, the earliest turbulence simulations were carried out before the IEEE standardization and used both higher and lower floating-point formats (e.g., early channel flow simulations used lower precision than today). Also, today some of the largest simulations of homogeneous isotropic turbulence are carried out in IEEE single precision (Yeung *et al.* 2015) due to lower memory requirements, and the smaller impact on the simulation result versus other sources of uncertainty. Single-precision simulations have also shown promising results for other types of simulations, such as implicit large-eddy simulations (Witherden & Jameson 2020), and for conventional finite-volume codes (Brogi *et al.* 2024). Lower precision has also garnered recent interest in the application area of weather and climate simulations, as covered by Palmer (2020).

This work focuses on assessing the physical limiting factors for numerical precision. While the methods employed, such as iterative solvers, can be hugely sensitive to round-off errors due to large global reduction, the goal of this study is not to assess how such a limitation can be overcome through mixed-precision linear algebra (which has also gathered increasing interest (Abdelfattah *et al.* 2021)). Instead, the aim is to understand the extent to which different terms in the non-linear dynamical system can be rounded to a lower floating-point format while still preserving the dynamics of the flow, and also whether the state of the flow itself can be represented with lower floating-point numbers.

However, due to the nonlinear nature of the Navier–Stokes equations, cases where numerical precision will have a larger-than-expected impact can be found. In particular, if there exist several attractors of the flow or a symmetry that is sensitive to small disturbances, lower floating-point precision can be detrimental to the validity of the simulation (Fleischmann *et al.* 2019; Qin & Liao 2022). Nevertheless, for many simulations, as noted before, floating-point precision does not seem to have a significant impact. This is especially true when turbulent flows are considered with focus on a statistical description, or for deterministic cases, such as transition, where an initial disturbance higher than the numerical noise is introduced. Also, it is vital to understand how numerical precision impacts the comparison to experimental data. Free-stream turbulence levels greater than 0.1% in experimental setups are common, and can be even larger in industrial applications. Therefore, the inherent uncertainty of a simulation due to small disturbances and unknown boundary conditions (compared with the actual physical setup) is significant. The question is to what extent and through what mechanism the numerical precision of a simulation increases this uncertainty or provides a biased result.

Name	bits	$b$	$b_e$	$\varepsilon = 2^{-b-1}$	Name	bits	$b$	$b_e$	$\varepsilon = 2^{-b-1}$
<b>FP64</b>	64	52	11	$2^{-53} \approx 2 \cdot 10^{-16}$	<b>bfloat16</b>	16	7	8	$2^{-8} \approx 4 \cdot 10^{-3}$
<b>FP32</b>	32	23	8	$2^{-24} \approx 6 \cdot 10^{-8}$	<b>E4M3</b>	8	3	4	$2^{-4} = 0.0625$
<b>FP16</b>	16	10	5	$2^{-11} \approx 5 \cdot 10^{-4}$	<b>E5M2</b>	8	2	5	$2^{-3} = 0.125$

TABLE 1. Different floating-point formats. The rounding machine epsilon  $\varepsilon$ ,  $|u_{\text{FP}} - u| < \varepsilon u$  for some real number  $u$ , is the largest round-off error introduced due to floating-point precision.

### 2.1. Floating-point numbers

A floating-point number is defined by a number of mantissa bits  $b$  and exponent bits  $b_e$  together with one sign bit  $s$  dictating the sign of the floating-point number. If we let  $e$  be the value of the exponent (as an unsigned  $b_e$ -bit integer), and  $c_i$  be the  $i$ th least significant bit of the mantissa, the value for a given normal floating-point number is  $(-1)^s \left(1 + \sum_{i=1}^b c_{b-i} 2^{-i}\right) \times 2^{e-(2^{b_e-1}-1)}$ . Floating-point numbers with deterministic rounding are the most readily available and commonly used in modern computing systems, and in this work we limit ourselves to this type of quantization. We consider floating-point numbers between 8 and 64 bits as listed in Table 1. In our work, floating-point numbers below FP32 are emulated with CPFloat (Fasi & Mikaitis 2023).

### 2.2. Perturbing the integration of the Navier–Stokes equations

To assess the impact of numerical precision, we perturb the simulation by rounding parts of it or by performing the entire simulation at lower precision. Across different solvers, a simulation can be described as a discrete map  $\mathbf{u}^{i+1} = f(\mathbf{u}^i)$ , where  $\mathbf{u}$  is the state at any given point in time. For incompressible flow, the state would most often be the velocity components and the pressure. To evaluate the impact of the perturbations introduced by different floating-point formats, we perform the rounding in three different ways:

(a) **Full FP32.** The entire solver is run using IEEE single precision. This is the only case where the lower precision is not emulated in this work.

(b) **State rounding.** Casting  $\mathbf{u}^i$  in lower precision, while the solver operates in FP64. Introducing the rounding operator  $\tilde{\cdot}$  corresponds to  $\tilde{\mathbf{u}}^{i+1} = f(\tilde{\mathbf{u}}^i)$ .

(c) **Term rounding.** Different terms in the Navier–Stokes equations, such as the convective or viscous term or both, are represented in lower precision; in other words, for the convective term it would be computed as  $u_j \widetilde{\partial u_i / \partial x_j}$  with the rounding operator.

In order to assess how these perturbations impact different numerical schemes and different formulations of the Navier–Stokes equations, we consider four different flow solvers with different discretizations and characteristics.

### 2.3. Software and numerical methods

#### 2.3.1. Neko

Neko is based on a continuous Galerkin spectral-element framework with a special focus on the incompressible Navier–Stokes equations, with extensive support for heterogeneous computer architectures (Jansson *et al.* 2024). The code has excellent scaling demonstrated to thousands of GPUs and was nominated for the Gordon Bell Prize in 2023. The solver

uses high-order hexahedral spectral elements (polynomial order 7 for the tests here), with the  $P_N - P_N$  method for velocity–pressure decoupling, a third-order semi-implicit time integration method, and dealiasing of the convective term using the 3/2-rule. The following tests are performed with Neko: Full FP32, perturbation of the convective term (denoted Convection FPX for precision FPX), and state rounding (denoted State FPX).

### 2.3.2. *Simson*

Simson is a fully spectral code for channel and boundary-layer geometries, based on Fourier discretization in the streamwise and spanwise directions, and Chebyshev expansion in the vertical (wall-normal) direction. The mesh is equidistant in the wall-parallel directions, and follows a Gauss–Lobatto distribution in the wall-normal direction. Standard dealiasing using the 3/2 rule is performed in the Fourier directions only. All solvers are direct in velocity–vorticity formulation; thus, no tolerances need to be specified. A comprehensive user guide is available in Chevalier *et al.* (2007). Tests performed using Simson include Full FP32, Convection FPX, and State FPX.

### 2.3.3. *SSDC*

SSDC implements a high-order entropy-stable discontinuous collocated Galerkin method for the compressible Navier–Stokes equations (Parsani *et al.* 2021). The mesh consists of hexahedral elements with support for an unstructured topology and nodes inside each element are distributed according to the Gauss–Legendre–Lobatto quadrature points. The solver has demonstrated good strong parallel scaling up to at least 100 000 CPU cores. The solver is explicit in time and the rounding is applied to the right-hand-side vector, at every stage of a Runge–Kutta type time integrator. The rounding is applied separately to the convective and viscous terms, and in three ways: (i) To the state vector (State Convective/Viscous FPX); (ii) To the flux terms before adding to the right-hand side (Convective/Viscous FPX); (iii) Performing both operations (Combined FPX).

### 2.3.4. *PadeLibs*

PadeLibs is a Navier–Stokes solver for high-resolution simulations of compressible turbulent flows (Song *et al.* (2024)). The numerical discretization uses sixth-order compact finite-difference methods with collocated variable storage and staggered flux assembly. The simulation framework used in PadeLibs is robust to aliasing errors and has high accuracy in resolving diffusive fluxes at small scales. In this work, round-off effects are investigated by rounding the convective (inviscid) fluxes to a precision FPX (Convective FPX) after they are assembled before taking the divergence operations. The rounded results still keep the double-precision format (FP64) although the emulated round-off errors are introduced. All the differential and interpolation operations are consistently calculated in double-precision format. The operator coefficients are all at double-precision accuracy, and the round-off errors are added only from the input. For the incompressible test cases, the Mach number is set to be 0.25.

## 3. Results

### 3.1. *Turbulent channel flow*

The first test case is turbulent channel flow at  $Re_\tau = 550$  in a relatively modest domain of  $2\pi\delta \times 2\delta \times \pi\delta$ . The resolutions follow standard practice for high-order simulations of wall turbulence:  $\Delta x^+ \approx 12$ ,  $\Delta z^+ \approx 5$ , and  $\Delta y^+$  similar to, for example, Del Alamo

Setup	$Re_\tau$	Avg. time	Domain size	$\max\% \frac{\langle u \rangle - \langle u \rangle_{\text{ref}}}{\langle u \rangle_{\text{ref}}}$	$\max\% \frac{\langle u' u' \rangle - \langle u' u' \rangle_{\text{ref}}}{\langle u' u' \rangle_{\text{ref}}}$
<b>Neko</b>					
Full FP64	548	$6.6\delta/u_\tau$	$2\pi\delta \times 2\delta \times \pi\delta$	0.64	6.1
Full FP32	554	$7.7\delta/u_\tau$	$2\pi\delta \times 2\delta \times \pi\delta$	3.7	4.6
Convective FP32	548	$5.5\delta/u_\tau$	$2\pi\delta \times 2\delta \times \pi\delta$	0.30	3.2
Convective FP16	552	$5.5\delta/u_\tau$	$2\pi\delta \times 2\delta \times \pi\delta$	1.0	2.0
Convective E5M2	551	$11.1\delta/u_\tau$	$2\pi\delta \times 2\delta \times \pi\delta$	1.25	7.7
Convective E4M3	553	$11.1\delta/u_\tau$	$2\pi\delta \times 2\delta \times \pi\delta$	1.6	10
State FP32	553	$5.5\delta/u_\tau$	$2\pi\delta \times 2\delta \times \pi\delta$	1.3	1.9
State FP16	683	$11.1\delta/u_\tau$	$2\pi\delta \times 2\delta \times \pi\delta$	17	37
<b>Simson</b>					
Full FP64	543	$86.9\delta/u_\tau$	$2\pi\delta \times 2\delta \times \pi\delta$	0.65	5.3
Full FP32	544	$43.4\delta/u_\tau$	$2\pi\delta \times 2\delta \times \pi\delta$	0.52	4.7
Full FP64 Coarse in $x$ & $z$	542	$43.5\delta/u_\tau$	$2\pi\delta \times 2\delta \times \pi\delta$	0.6	3.5
Full FP32 Coarse in $x$ & $z$	544	$43.5\delta/u_\tau$	$2\pi\delta \times 2\delta \times \pi\delta$	0.36	4.5
Full FP64 Large Domain	544	$145\delta/u_\tau$	$8\pi\delta \times 2\delta \times 3\pi\delta$	0.14	1.0
Full FP32 Large Domain	543	$145\delta/u_\tau$	$8\pi\delta \times 2\delta \times 3\pi\delta$	0.10	1.6
Full FP32 Large Domain	1000	$92\delta/u_\tau$	$8\pi\delta \times 2\delta \times 3\pi\delta$	0.75	2.0
<b>SSDC</b>					
Full FP64	546	$10.93\delta/u_\tau$	$2\pi\delta \times 2\delta \times \pi\delta$	0.42	7.2
State FP32	548	$10.95\delta/u_\tau$	$2\pi\delta \times 2\delta \times \pi\delta$	0.54	11
State FP16	550	$11.00\delta/u_\tau$	$2\pi\delta \times 2\delta \times \pi\delta$	1.0	20
State Convective FP32	548	$10.95\delta/u_\tau$	$2\pi\delta \times 2\delta \times \pi\delta$	0.54	11
State Convective FP16	550	$11.00\delta/u_\tau$	$2\pi\delta \times 2\delta \times \pi\delta$	1.1	20
State Viscous FP32	547	$10.95\delta/u_\tau$	$2\pi\delta \times 2\delta \times \pi\delta$	0.54	5.9
State Viscous FP16	545	$10.91\delta/u_\tau$	$2\pi\delta \times 2\delta \times \pi\delta$	1.1	4.5
<b>PadeLibs</b>					
Full FP64	565	$4.6\delta/u_\tau$	$2\pi\delta \times 2\delta \times \pi\delta$	4.1	9.6
Convective FP32	565	$3.5\delta/u_\tau$	$2\pi\delta \times 2\delta \times \pi\delta$	4.1	9.6
Convective FP16	565	$3.3\delta/u_\tau$	$2\pi\delta \times 2\delta \times \pi\delta$	4.1	9.6
Convective E5M3	565	$4.1\delta/u_\tau$	$2\pi\delta \times 2\delta \times \pi\delta$	4.1	9.6

TABLE 2. Details for the different channel-flow simulations. Additional tests were carried out using Simson, including the effect of Reynolds number, resolution, domain size, and time step. The reported error values might be impacted by relatively modest averaging times. The compressible codes (SSDC and PadeLibs) are expected to have higher error levels compared to reference data (Lee & Moser 2015) due to compressibility effects.

*et al.* (2004). Table 2 summarizes the simulation parameters of the runs that did not diverge. The maximum relative difference of the first- and second-order moments for the streamwise velocity  $u$  are compared with the reference data from Lee & Moser (2015). Overall, the results of the different simulations were largely unaffected by low precision down to FP16. Especially for first-order moments the solution is not visibly sensitive. However, for second-order moments, and specifically the streamwise fluctuations, lower precision than FP32 can have detrimental effects.

It is difficult to discern the impact on statistics of the numerical precision from the domain size and time averaging. This motivated a consideration of larger domains, longer

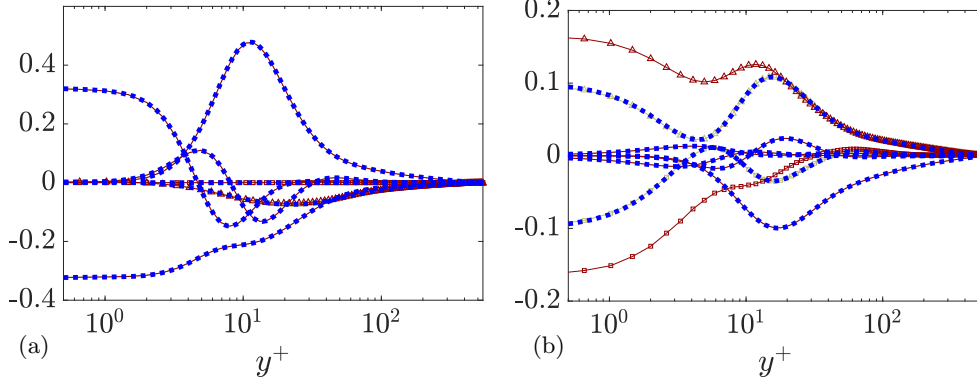


FIGURE 1. Budget terms in the transport equation of  $\langle u'u' \rangle$  (a) and  $\langle u'v' \rangle$  (b) for the turbulent channel flow at  $Re_\tau \approx 550$  using FP64 (beige) and FP32 (dark red) compared to the reference data from Lee & Moser (2015) (dotted blue). Results are computed using Simson. Resolutions are identical between FP32 and FP64. Triangles and squares denote the pressure-strain and pressure transport terms, respectively.

averaging times, and higher Reynolds numbers, especially to confirm the validity of the Full FP32 simulations. Overall, we concluded that the differences observed previously were primarily due to domain size and averaging time, as an increase in both reduced the relative error to 0.1%–0.2%. The resulting profiles then showed excellent agreement with the reference data. In addition, we observed that the impact of numerical precision was not sensitive to the number of Fourier modes, for instance, due to a larger domain size (see the Large Domain runs in Table 2). We also tested performing only the global transpose communications in FP32 (not shown) while keeping all computation in FP64, with indistinguishable results from FP64.

A more involved analysis of budget terms in the transport equation was performed with Simson. The components related to pressure-velocity coupling (pressure-strain and pressure transport) were the only ones sensitive to arithmetic precision at FP32, as shown in Figure 1. For the higher  $Re_\tau \approx 1000$ , we observed a nearly zero value for all pressure-related terms of all Reynolds stress components, suggesting a complete decorrelation of the instantaneous pressure field from the instantaneous velocity and its derivative. However, given the velocity–vorticity formulation in Simson, the instantaneous pressure does not enter the evolution of the flow, and it is computed as a separate step only if needed. Since the budget terms related to velocity gradients are robust to precision, the observed differences are likely caused by a sensitivity to precision due to the specific implementation of the Poisson solver, which can be addressed by making minor changes to the code. Further investigation is needed in this regard.

### 3.2. Transition to turbulence

We consider K-type transition where a laminar baseflow as described by Schlatter (2005) is perturbed by a 2D and two oblique 3D Tollmien–Schlichting (TS) waves with amplitudes of 3% and 0.1% respectively (based on the centerline velocity), all of which are individually stable. The Reynolds number is  $Re_b = 3333$  based on the constant bulk velocity, which corresponds to  $Re_{cl} = 5000$  of the initial parabolic velocity profile. The domain size is  $5.61\delta \times 2.99\delta \times 2\delta$ , adjusted to fit the chosen TS waves. For the compressible codes, a constant forcing is applied in the streamwise direction to drive the flow instead

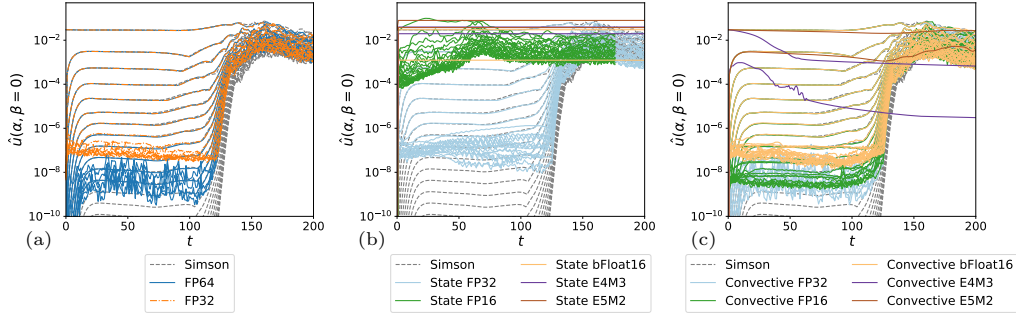


FIGURE 2. Evolution in time of amplitude of 2D modes  $|\hat{u}(\alpha, \beta = 0)|$  for transitional case. All simulations carried out in Neko, except the reference case in Simson shown in gray. Results from full FP32 and FP64 (a), State rounding (b), and rounding of the convective term (c) are shown.

of fixed bulk velocity. In all cases, a matching bulk Reynolds number of  $Re_b = 3333$  is maintained before turbulent breakdown.

The evolution of selected Fourier modes is shown in Figure 2. The 2D modes  $|\hat{u}(\alpha, \beta = 0)|$  show the quick establishment of a saturated 2D TS wave, with a weak temporal decay. Secondary instability, initiated by the  $\beta = 1$  modes, leads to a quick increase in the energy in all modes ( $t > 120$ ) and subsequent breakdown to turbulence ( $t > 175$ ). The double-precision arithmetic allows us to resolve numerically all modes down to machine precision ( $10^{-15}$ ) for Simson, but saturates at around  $10^{-9}$  for the other solvers. Reducing to single precision increases the ambient noise level to about  $10^{-8}$  for Simson, and around one order of magnitude higher for the other codes. Interestingly, there seems to be no interaction between these modes that would lead to a premature growth in the physically relevant modes. In contrast, similar studies using low-resolution simulations found a clear change of energy distribution and subsequent growth, which can be contained only by using appropriate subgrid-scale models (Schlatter 2005). From Figure 2 we can conclude that the evolution of the individual modes, but also integral quantities such as the global friction or centerline velocities, are not dependent on the precision. For the rounding of the state and convective terms in the different solvers, we also observe that FP32 performs remarkably well, but when representing the state at lower-precision, the simulation becomes prone to stagnation (horizontal lines) or an immediate transition (State FP16). However, although the transitional case is sensitive, the amplitude of the initial conditions is still on the order of 0.1–1%, and there is likely a precision-dependent limit on the smallest disturbance amplitude the simulations would be able to capture. In addition, the geometry is still a Cartesian channel, which motivates the study of a deformed geometry, such as the separating flow around a cylinder.

### 3.3. Cylinder $Re_D = 3900$

This section considers the flow around an infinite circular cylinder at  $Re_D = U_\infty D / \nu = 3900$ , where  $D$  is the cylinder diameter and  $U_\infty$  the free-stream velocity. We perform LES with approximately 512 grid points along the cylinder boundary, and a spanwise length of  $2\pi D$  with 128 grid points. There is extensive literature on this case, showing a significant spread in the simulation results (Lehmkuhl *et al.* 2013). The results are illustrated in Figure 3, which shows the velocity profiles in the wake and pressure distribution on the cylinder surface. Table 3 compares these results with the original LES by Kravchenko & Moin (2000) and highlights wall quantities such as the drag coefficient and separation

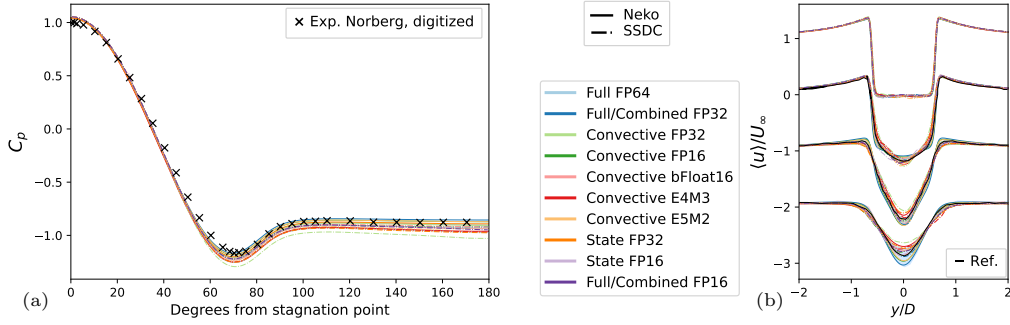


FIGURE 3. Profiles for the cylinder at  $Re_D = 3900$  in Neko and SSDC with rounding the convective term, the state, and running the entire solver in single and double precision. The  $C_p$  profile in the center of wake (a) and wake profile at four different locations in the wake (0.58, 1.06, 1.54, 2.02) (b), the blue-shaded interval is between time averages of two low-frequency modes as described by Lehmkuhl *et al.* (2013).

Setup	Avg. time ( $D/U_\infty$ )	$f_{vs}$	$\phi_s$	$L_r$	$\bar{C}_d$	$-\bar{C}_{Pb}$
Kravchenko & Moin (2000)	35	0.21	88	1.35	1.04	0.94
Neko						
Full FP64	300.3	0.2097	86.62	1.48	0.9926	0.9159
Full FP32	319.1	0.2068	86.46	1.553	0.9911	0.8979
Convective FP32	100.0	0.2087	86.97	1.371	1.025	0.9496
Convective FP16	100.0	0.2087	86.62	1.481	1.004	0.9149
Convective bFloat16	100.0	0.2087	87.18	1.271	1.042	0.9775
Convective E4M3	73.13	0.204	87.49	1.182	1.057	1.003
Convective E5M2	77.46	0.2054	86.64	1.472	1.003	0.9148
State FP32	100.0	0.2087	86.8	1.402	1.016	0.9347
SSDC						
Full FP64	300.0	0.2075	87.1	1.336	1.076	0.9731
Combined FP32	100.0	0.2035	86.7	1.391	1.058	0.9185
Combined FP16	100.0	0.2050	86.6	1.386	1.062	0.9475
Convective FP32	100.0	0.2050	87.5	1.149	1.117	1.0285
Convective FP16	100.0	0.2050	86.8	1.272	1.061	0.9233
Convective bfloat16	100.0	0.2050	86.6	1.405	1.054	0.9202
Convective E4M3	30.0	0.2099	87.1	1.205	1.069	0.9695
Convective E5M2	100.0	0.2064	86.8	1.272	1.093	0.9325
State FP32	100.0	0.2035	87.2	1.183	1.085	0.9756
State FP16	100.0	0.2099	86.8	1.386	1.063	0.9305
PadeLibs						
Full FP64	69.0	0.2093	87.35	1.348	0.9932	0.9596
Convective FP16	36.6	0.2093	87.35	1.297	0.9932	0.9596
Convective E5M2	56.3	0.2097	88.86	1.028	1.073	1.0352

TABLE 3. Scalar values associated with the cylinder at  $Re_D = 3900$ . Columns correspond to each setup name, the time statistics were collected for, the separation angle  $\phi_s$ , the recirculation length  $L_r$ , the drag coefficient  $\bar{C}_d$ , and the base pressure coefficient  $\bar{C}_{Pb}$ .

angle, as well as the length of the recirculation zone. Overall, the differences among the setups and solvers are comparable to the spread in the reference data. As such, for the simulations that do not diverge, this case indicates that other sources of uncertainty are more significant than the numerical precision when it comes to LES of separating flows. This includes the size of the domain and, in particular, averaging times. Isolating the impact of lower precision might become clearer with longer averaging times, but due to the discrepancies among multiple reference data it is not certain whether the impact of precision can be isolated.

#### 4. Conclusions

We have carried out simulations of three flow cases using four different flow solvers, introducing lower-precision floating-point numbers to different terms in the equations, as well as executing entire simulations in FP32. Overall, the results indicate that numerical precision can be significantly lower for simulations of turbulence, transition, and separation than conventional FP64 without causing relevant deterioration of simulation results, much to our surprise, especially when considering transition. In particular, single-precision simulations open an avenue for significantly lower computational costs with only marginal reduction in quality. While single-precision arithmetic might not work out of the box for all solvers, and puts additional strain on the developer, the cost reductions are about a factor of two. Considering that large-scale CFD simulations utilize many millions of core hours, it appears that all codes should apply significant effort to utilize lower precision to both save monetary costs and free up storage. In addition, lower-precision arithmetic allows to compute the results faster and to exploit consumer-grade GPUs that support lower precision formats and are available at a significantly lower cost than server-grade components. Although for some numerical algorithms the requirements with regard to floating-point numbers might differ, our results illustrate that the overarching flow physics is accurately represented even below FP64. Going forward, we intend to increasingly utilize lower floating-point formats and continue the evaluation of the flow physics to test the impact of precision for simulations of turbulence.

#### 5. Acknowledgements

The computations were enabled by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council (grant 2022-06725), and project access to the LUMI supercomputer, owned by the EuroHPC Joint Undertaking and hosted by CSC (Finland) and the LUMI consortium. We acknowledge the HPC resources provided by the Erlangen National High Performance Computing Center (NHR@FAU) of the FAU Erlangen-Nürnberg for project b237dc. NHR@FAU hardware is partially funded by the German Research Foundation (DFG) – 440719683. This research also used Shaheen III managed by the Supercomputing Core Laboratory at King Abdullah University of Science & Technology (KAUST) in Thuwal, Saudi Arabia.

#### REFERENCES

- ABDELFATTAH, A., ANZT, H., BOMAN, E. G., CARSON, E., COJEAN, T., DONGARRA, J., FOX, A., GATES, M., HIGHAM, N. J., LI, X. S. *et al.* 2021 A survey of numerical linear algebra methods utilizing mixed-precision arithmetic. *Int. J. High Perform. Comput. Appl.* **35**, 344–369.

- BROGI, F., BNÀ, S., BOGA, G., AMATI, G., ONGARO, T. E. & CERMINARA, M. 2024 On floating point precision in computational fluid dynamics using OpenFOAM. *Future Gener. Comput. Syst.* **152**, 1–16.
- CHEVALIER, M., SCHLATTER, P., LUNDBLADH, A. & HENNINGSON, D. S. 2007 SIMSON—a pseudo-spectral solver for incompressible boundary layer flows. *Tech. Rep. TRITA-MEK 2007:07*. KTH Mechanics, Stockholm, Sweden.
- DEL ALAMO, J. C., JIMENEZ, J., ZANDONADE, P. & MOSER, R. D. 2004 Scaling of the energy spectra of turbulent channels. *J. Fluid Mech.* **500**, 135–144.
- FASI, M. & MIKAITIS, M. 2023 Cpfloat: A library for simulating low-precision arithmetic. *ACM T.Math. Software* **49**, 1–32.
- FLEISCHMANN, N., ADAMI, S. & ADAMS, N. A. 2019 Numerical symmetry-preserving techniques for low-dissipation shock-capturing schemes. *Comput. Fluids* **189**, 94–107.
- HOMANN, H., DREHER, J. & GRAUER, R. 2007 Impact of the floating-point precision and interpolation scheme on the results of DNS of turbulence by pseudo-spectral codes. *Comput. Phys. Commun.* **177**, 560–565.
- JANSSON, N., KARP, M., PODOBAS, A., MARKIDIS, S. & SCHLATTER, P. 2024 Neko: A modern, portable, and scalable framework for high-fidelity computational fluid dynamics. *Comput. Fluids* **275**, 106243.
- KRAVCHENKO, A. G. & MOIN, P. 2000 Numerical studies of flow over a circular cylinder at  $Re_D = 3900$ . *Phys. Fluids* **12**, 403–417.
- LEE, M. & MOSER, R. D. 2015 Direct numerical simulation of turbulent channel flow up to  $Re_\tau \approx 5200$ . *J. Fluid Mech.* **774**, 395–415.
- LEHMKUHL, O., RODRÍGUEZ, I., BORRELL, R. & OLIVA, A. 2013 Low-frequency unsteadiness in the vortex formation region of a circular cylinder. *Phys. Fluids* **25**, 085109.
- PALMER, T. 2020 Number formats, error mitigation, and scope for 16-bit arithmetics in weather and climate modeling analyzed with a shallow water model. *J. of Adv. in Model. Earth Sy.* **12**, e2020MS002246.
- PARSANI, M., BOUKHARFANE, R., NOLASCO, I. R., DEL REY FERNÁNDEZ, D. C., ZAMPINI, S., HADRI, B. & DALCIN, L. 2021 High-order accurate entropy-stable discontinuous collocated Galerkin methods with the summation-by-parts property for compressible CFD frameworks: Scalable SSDC algorithms and flow solver. *J. Comput. Phys.* **424**, 109844.
- QIN, S. & LIAO, S. 2022 Large-scale influence of numerical noises as artificial stochastic disturbances on a sustained turbulence. *J. Fluid Mech.* **948**, A7.
- REZAEIRAVESH, S., VINUESA, R. & SCHLATTER, P. 2021 On numerical uncertainties in scale-resolving simulations of canonical wall turbulence. *Comput. Fluids* **227**, 105024.
- SCHLATTER, P. 2005 Large-eddy simulation of transition and turbulence in wall-bounded shear flow. PhD thesis, ETH Zürich, Switzerland.
- SONG, H., GHATE, A. S., MATSUNO, K. V., WEST, J. R., SUBRAMANIAM, A. & LELE, S. K. 2024 A robust compact finite difference framework for simulations of compressible turbulent flows. *J. Comput. Phys.* **519**, 113419.
- WITHERDEN, F. & JAMESON, A. 2020 Impact of number representation for high-order implicit large-eddy simulations. *AIAA J.* **58**, 184–197.
- YEUNG, P., ZHAI, X. & SREENIVASAN, K. R. 2015 Extreme events in computational turbulence. *P. Natl. Acad. Sci. USA* **112**, 12633–12638.