Mixed-Precision Computing in the GRIST Dynamical Core for Weather and Climate Modelling

Siyuan Chen[1,5], Yi Zhang[1,3,5], Yiming Wang[1,5], Zhuang Liu[2], Xiaohan Li[2], Wei Xue[4]

1 2035 Future Laboratory, PIESAT Information Technology Co., Ltd., Beijing, China

5 2 Ministry of Education Key Laboratory for Earth System Modelling, Department of Earth System Science, Tsinghua University, Beijing, China

3 Chinese Academy of Meteorological Sciences, Beijing, China

4 Department of Computer Science and Technology, Tsinghua University, Beijing, China

5 Beijing Research Institute, Nanjing University of Information Science and Technology, Beijing, China

Correspondence to: Yi Zhang (zhangyi_fz@piesat.cn)

10 **Abstract.** Atmosphere modelling applications become increasingly memory-bound due to the inconsistent development rates between processor speeds and memory bandwidth. In this study, we mitigate memory bottlenecks and reduce the computational load of the GRIST dynamical core by adopting the mixed-precision computing strategy. Guided by a limited-degree of iterative development principle, we identify the equation terms that are precision insensitive and modify them from double- to single-precision. The results show that most precision-sensitive terms are predominantly linked to pressure-gradient and gravity terms, while most precision-insensitive terms are advective terms. The computational cost is reduced without compromising the solver accuracy. The runtime of the model's hydrostatic solver, non-hydrostatic solver, and tracer transport solver is reduced by 24%, 27%, and 44%, respectively. A series of idealized tests, real-world weather and climate modelling tests, has been performed to assess the optimized model performance qualitatively and quantitatively. In particular, in the high-resolution weather forecast simulation, the model sensitivity to the precision level is mainly dominated by the small-scale features. While in long-term climate simulation, the precision-induced sensitivity can form at the large scale.

1 Introduction

Increasing model resolution is an effective approach of enhancing the atmosphere model forecast accuracy (Bauer et al. 2021; Benjamin et al.2019; Yu et al. 2019). Highly accurate, efficient, stable and scalable global dynamical cores have been widely pursued over the past two decades (e.g., Tomita and Satoh 2004; Harris and Lin 2012; Skamarock et al. 2012; Zängl 25 et al. 2015; Wedi et al. 2020; Sergeev et al. 2023; Zhang et al. 2023). Doubling the horizontal resolution with a fixed vertical resolution leads to an increase of computational amount by a factor of $\sim 2^3$, a significant challenge in terms of computational cost and energy consumption.

Operational weather and climate forecasting is a field where the dual demands of accuracy and computational efficiency converge, necessitating both quality and speed. In the context of high-resolution meso-scale forecasting, which operates on

30 scales of a few kilometers, computational efficiency itself implies forecast accuracy. Faster models enable more frequent forecast-assimilation cycles and the use of larger ensemble sizes within the constraints of finite computational resources. To tackle these computational hurdles, efforts have concentrated on enhancing the efficiency of numerical models. Progress such as Field Programmable Gate Arrays (FPGAs) and heterogeneous computing (e.g., Gan et al. 2013; Yang et al. 2016; Fu et al. 2017; Gu et al. 2022; Taylor et al. 2023), alongside compiler optimizations (e.g., Santos et al. 2024), have

35 demonstrated significant potential in accelerating earth system model.

Conventional weather/climate model development has typically relied on double-precision (64-bit) floating-point. The transition from double- to single-precision (32-bit) or even half-precision floating-point arithmetic presents an intriguing avenue for enhancing the computational efficiency (Düben et al. 2014). Single-precision computation unveils several compelling advantages, especially when confronted with the memory wall (Abdelfattah et al. 2021; Fornaciari et al. 2023;

40 Brogi et al. 2024). Beyond the alleviation of memory constraints, single-precision arithmetic promises three distinct benefits: accelerated arithmetic operations, improved cache hit rates, and reduced inter-node data communication (Baboulin et al. 2009; Düben and Palmer 2014; Düben et al. 2015; Váňa et al. 2016; Nakano et al. 2018). The benefits highlighted illustrate the capability of single-precision computation to boost computational efficiency in high-performance computing tasks, especially within the realm of large-scale weather and climate simulations where computational expenses are significant.

45 However, a wholesale migration from double- to single-precision computing may not always yield beneficial outcomes. This has led to the exploration of precision sensitive model component and/or physical scale in earth system modeling (e.g., Thornes et al. 2017; Nakano et al. 2018; Chantry et al. 2019; Maynard and Walters 2019; Cotronei and Slawig 2020). Single-precision algorithms may struggle to converge or achieve the required precision when tackling intricate fluid dynamics simulations. In certain scenarios, single-precision computations can also result in floating-point under/overflow (Váňa et al.

50 2016; Cotronei et al. 2020). Additionally, physical parameterization schemes in the atmospheric models may amplify the grid-scale oscillations when executed in a pure single-precision mode (Váňa et al. 2016). Therefore, it becomes imperative to identify the specific algorithms within the modeling framework that are sensitive to the precision level.

Previous studies have made a notable progress. A pivotal study (Váňa et al. 2016) explored the reduction of almost all real-number variables in the Integrated Forecast System (IFS) of the European Center for Medium-Range Weather

55 Forecasting (ECMWF) from 64 bits to 32 bits. Results revealed that reducing precision did not significantly compromise the model accuracy, while it considerably reduced the computational burden by a factor of ~40%. Based on the dynamical core of the nonhydrostatic icosahedral model (NICAM), Nakano et al. (2018) witnessed an undesirable wavenumber-5 structure when completely using single-precision computing. This abnormal wave growth was traced back to the errors in the grid cell geometry calculations. By using double precision for only necessary parts in the dynamical core and single precision for all

60 other parts, the model successfully simulated the baroclinic wave growth, and achieved a ~46% reduction of runtime. Based on the Yin-He global spectral model, Yin et al. (2021) used a single-precision fast spherical harmonic transform (SHT) to conduct a 10-day global simulation and a 30-day retrospective forecasting experiment. Their simulations reproduced the

major precipitation events over southeastern China. The single-precision fast SHT may lead to a reduction of runtime by ~25.28% without significantly affecting the forecasting skill. Cotronei et al. (2020) converted the majority of the computations within the radiation component of European Centre Hamburg Model (ECHAM) to single-precision, resulting in a 40% reduction in the runtime of the individual component. The obtained results were comparable to those achieved with double-precision. Banderier et al. (2023) indicated that employing single precision for regional climate simulations can significantly reduce computational costs (~30%) without significantly compromising the quality of model results.

While these studies have demonstrated various ways for precision optimization, certain limitations remain. First, some studies focused on a complete transition to single-precision, potentially overlooking the precision-sensitive components and lacked a discussion of optimization strategies. Moreover, the applicability of mixed-precision in global climate simulations remains to be validated. Furthermore, because of the diversity of numerical models and algorithms, encompassing grid systems and solver techniques, these differences may lead to the model-specific precision sensitivity. Certain algorithms may remain amenable to single-precision computations, while others necessitate the use of double precision for stability and accuracy. These gaps in the literature underscore the need for the present research to explore precision sensitivity, and to test the reduced-precision computing for both weather and climate simulations.

In this study, we explored the strategies of mixed-precision computing in the dynamical core of the Global-Regional Integrated Forecast System (GRIST; Zhang et al. 2019; Zhang et al. 2020). GRIST is a unified weather-climate model system designed for both research and operational modeling applications. Through a detailed implementation by modifying certain parts of the original (double-precision) dynamical core to support single-precision, a significant reduction of the computational burden has been achieved without sacrificing the solution accuracy, stability, and physical performance. This has been validated based on a series numerical tests ranging from idealized to real-world flow.

The remainder of this paper is organized as follows. Section 2 introduces the GRIST model, presents the mixed-precision optimization strategies, code modifications and highlighting the key equation terms sensitive to precision. Section 3 examines the computational performance of mixed-precision computing. Section 4 evaluates the physical performance of mixed-precision computing in a series of test cases. Discussion and conclusion are given in Sections 5 and 6, respectively.

2.1. GRIST

The GRIST dynamical core employs layer-averaged governing equations based on the generalized hybrid sigma-mass vertical coordinate and a horizontal unstructured grid, allowing a switch between the hydrostatic and non-hydrostatic solvers (Zhang 2018; Zhang et al. 2019; Zhang et al. 2020). Prognostic variables are arranged in a hexagonal Arakawa-C grid approach. The hydrostatic solver is fully explicit, based on the Runge-Kutta integrator and the Mesinger forward-backward scheme. The nonhydrostatic solver employs a horizontally explicit vertically implicit approach. There is no time splitting in the integration of the dry dynamical core (dycore hereafter), while the tracer transport module is time-splitted from the dycore, and supports several transport schemes for various applications (Zhang et al. 2020). In this study, a third-order

95   upwind flux operator combined with the Flux-Corrected Transport limiter is used in the horizontal, and an adaptively implicit method is used in the vertical (cf., Li and Zhang 2022).

2.2 Mixed-precision optimization strategy

The purpose is to decrease the precision level (and thus computational cost) and maintain the accuracy and stability. Before implementing mixed-precision computing, we have checked that completely using single precision for the entire
100   dynamical core leads to an unacceptable accuracy loss (see Section 4.1). However, considering the extensive codebase and its degree of complexity, comprehensively and randomly testing every component and variable is impractical. An iterative development approach with a minimum degree of trial and error is used to identify the model components that are sensitive to the precision level. The dry baroclinic wave of Jablonowski and Williamson (2006) is used as a benchmark test during the iterative development cycle because this case has complex fluid dynamics characteristics and is very sensitive to numerical
105   precision.

We established an acceptable error threshold, α, to assess whether the difference between outcomes from double-precision and mixed-precision simulations falls within a tolerable limit. Results from original double-precision computing serve as the true values. The iteration is: initially, a 10-day simulation was executed. We then embarked on a series of precision reduction tests for selected model variables, and we computed the error norm of selected diagnostic variables for
110   each test. E is defined to represent the overall error level:

$$E = \max\big(L(\mathcal{H})\big) \tag{1}$$

$$L(x) = \max\big(L_1(\mathcal{H}), L_2(\mathcal{H}), L_\infty(\mathcal{H})\big) \tag{2}$$

where L_1, L_2, and L_∞ represent the first, second and infinite norm of variable $\mathcal{H}$. The definitions of L_1, L_2, and L_∞ can be found in the appendix. Should "error" E exceed α (0.05 for this study), the modification is deemed unacceptable and
115   consequently abandoned; otherwise, the modification is accepted, allowing a further reduction in variable precision based on this new configuration. The precision optimization tests were conducted using the G8 grid. The grid names and their corresponding resolutions can be found in Table 1. Initially, selected diagnostic variables ($\mathcal{H}$) are ps (surface pressure) and vor (relative vorticity), because they can effectively quantify deviations in the mass field and velocity field. This criterion was set beforehand, while it has turned out that $L(vor)$ overall has much larger error magnitude than $L(ps)$. Thus, it is
120    $L(vor)$ determines our optimization outcome.

Technically, the switch between double-precision and single-precision code is defined through the FORTRAN KIND parameter, specified in a constant module. As single-precision results may not always replicate the double-precision results and can occasionally generate unacceptable errors (e.g., see Section 3.2), it is crucial to identify precision-sensitive variables and solver components. An additional parameter 'ns' has been introduced in this constant module for the precision-
125   insensitive variables. This modification facilitates the transition between double-precision, single-precision, and mixed-precision computations. Please note that only the subroutine of the solver is modified, indicating that the model initialization section remains in double-precision operations. If the solver requires single-precision operands, double-precision variables

need to be converted to single-precision after initialization. This method ensures a streamlined transition to mixed precision with minimal changes to the code structure.

130   Some important aspects are summarized as follows:

(i)   Model variables insensitive to the precision level are set to the type parameter "ns". When "ns" is defined as single precision, the code executes mixed-precision computations; when defined as double precision, the code regresses double-precision computing and produces identical solutions as the original unmodified code;

(ii)   Appropriately decompose computations involving implicit type conversions to reduce performance degradation due

135   to precision conversion. For instance, 'a = b * c'. Here, 'a' is a single-precision floating point, 'b' a larger double-precision float, and 'c' a single-precision float. The conversion of 'c' to double-precision can introduce extra rounding errors. These errors, amplified by 'b', may accumulate over time, adversely affecting model outcomes. Single-precision calculations provide a consistent error boundary, unlike mixed-precision which introduces uncertainty. In some cases, results might even be better if the computation of a function were entirely in single

140   precision. Hence, optimization should proceed with caution, considering these error dynamics.

(iii)   The Message Passing Interface (MPI) communication was modified for single-precision variables; The built-in functions such as 'HUGE' or 'TINY' are used to obtain very large or very small values respectively, to ensure the values fall within the precision range of the variables.

2.3 Mixed-precision optimization results

145   Following the strategy outlined in Section 2.2, the mixed-precision GRIST dynamical core is established. The optimization results, as depicted on the left side of Fig. 1, are summarized based on the continuous-form governing equations. The meaning of each variable in the equations *exactly* follows Zhang et al. (2020) so that we avoid repeating explanation. Model variables with underlined text denote single-precision operands, variables in black represent double-precision operands. Black dashed boxes indicate that this part uses double-precision variables for computation, but the tendency is

150   saved as single precision. Gray shading indicates that this variable is diagnosed mostly from single-precision variables. Specifically, $\zeta_p = \frac{\zeta_a}{\delta\pi}$ is highly sensitive to the precision of $\delta\pi$, requiring a double precision $\delta\pi$.

     For the dycore, the precision sensitivity varies among different terms. The precision-sensitive terms are primarily related to pressure gradient and gravity terms. The precision-insensitive terms are mainly advective, which may tolerate lower numerical precision. Computationally, the advective parts of the equations are using higher-order operators which are

155   responsible for the major computational burden. The passive tracer transport equation (Eq. 10) can be mostly computed using single precision. The only part needs a careful modification is the black solid box, which indicates that it uses single-precision variables for computing, but the result is saved as double precision. $\delta\pi V$ (representing the mass flux) in Eq. (10) is accumulated and averaged from $\delta\pi V$ in Eq. (3), so computing it uses single precision. But when using it for tracer transport, this variable is converted to double precision so that the mass continuity equation of tracer transport uses a double-precision

160   mass flux.

The mass continuity equation Eq. (3) is solved using a flux form, ensuring global mass conservation of $\pi_s - \pi_t$ (π_t is a constant) within the bounds of machine rounding errors, which is at the double-precision level. Using single precision $\delta\pi V$ implies that mass continuity is locally conserved at the single precision level. Recognizing the potential importance of local mass conservation (e.g., Thuburn 2008), a compilation switch is designed, so that approximating $\delta\pi V$ and the related mass continuity tendency can be achieved in either single-precision or double-precision. The time difference between approximating the continuity equation using single-precision and double-precision accounts for ~1% -2% of the total computational time. We will examine the model sensitivity to this operation in Section 4.4.

2.4 Model code modification

Figure 1 illustrates the modification made to the original model code repository. Thanks to the modular structure of the model code, the mixed-precision version of the model dynamics can be seamlessly integrated as an add-on component, allowing for independent development. The switch between double-precision and mixed-precision dynamics is governed by the model's control unit, facilitating the transition between two code repositories via a compiler option ("MIXCODE"). Additional adjustments for each component include modifying parallel exchange functions to support reduced-precision variables, altering the precision level of allocated dynamics data, accommodating precision changes of specific variables in physics-dynamics coupling, and introducing a precision control variable. All these supplementary modifications (excluding those in the infrastructure layer) are also designated by the compiler option "MIXCODE". The pure single-precision code is achieved by simply using single-precision for all variables , marked as "SPCODE" in the code.

3 Computational performance

Before showing the physical performance, we first examine the computational performance of the optimized dynamical core. All computing performance are carried out on a local supercomputing cluster. Each computing node is equipped with 128GB memory, and the Central Processing Unit (CPU) is a Hygon C86 7285 model at 2.0 GHz. Each CPU features a 32 KB L1 data cache, a 64 KB L1 instruction cache, a 512 KB L2 cache, and an 8192 KB L3 cache. We use "SGL" to denote pure single precision computing, "DBL" to denote pure double precision, and "MIX" to represent mixed precision computing. All experiments were conducted on a G8 grid, submitted with the same topology: 756 MPI tasks distributed across six nodes.

Compared to the double-precision model, the runtime of the mixed-precision model for the non-hydrostatic dry dynamical core (NDC), hydrostatic dry dynamical core (HDC) and tracer transport solver reduced by 27%, 24% and 44% (Table 2). The runtime of the mixed-precision dycore solver is still larger compared to the single-precision dycore, indicating the time overhead incurred by the use of double-precision in precision-sensitive algorithms. The runtime of the mixed-precision tracer transport solver is comparable to that of the single-precision tracer transport solver, as most computations in the tracer transport module now use single-precision computing. It should be noted that the time gains from mixed-precision computing may also depend on hardware and compiler options (e.g., Brogi et al. 2024). This is out of the scope of this study.

4 Physical Performance

Due to the inherent complexity of atmosphere models, which have dycore, tracer transport, and model physics suites, directly assessing errors induced by precision in the real cases is not precise. To ensure robustness, a hierarchy of five test cases from simple to complex is adopted for model evaluation.

4.1 Moist baroclinic wave

This case is from the DCMIP2016, as outlined by Ullrich et al. (2014), a modified approach to the dry baroclinic instability scenario (Jablonowski and Williamson 2006). This experimental setup triggers the emergence of an unstable baroclinic wave pattern, initiated by early perturbations, which exhibits exponential growth and attains its maximum intensity around the 11th day. The experiment incorporates a passive tracer representing water vapor, which is subject to passive advection. Although the mixing ratio marginally influences the pressure gradient force, as noted by Zhang et al. (2020), the overall behaviour of wave growth is in substantial agreement with that in the dycore (Zhang et al. 2019). The primary objective is to assess the model's efficacy in replicating the typical dynamics of moist atmospheric conditions across various precision settings.

Figure 2 shows surface pressure and relative vorticity field at the model level near 850hPa (model layer 23) at day 11, as simulated by the G8 resolutions. The baroclinic waves shown the anticipated growth in the DBL simulation (Figs. 2a). In the SGL simulation, the primary growth fluctuations in the DBL simulation were reproduced (Figs. 2c). However, in the Northern Hemisphere, there were developments of incorrect spurious waves, whose intensity was comparable to the major fluctuations (Figs. 2c). The Southern Hemisphere exhibited a weaker structure of spurious waves (Figs. 2c). The results from the MIX simulation displayed patterns much closer to those in the DBL simulation (Figs. 2e).

The primary difference between MIX and DBL simulations lies in the vicinity of strong gradients along the cold front (Figs. 2c). But the primary fluctuations in both MIX and DBL simulations exhibit a high degree of similarity in their patterns (Figs. 2a and 2e), indicating that precision levels have a tangible impact on the phase speed of wave propagation.

The error introduced by SGL and MIX can be quantified by comparing solutions to a DBL solution. Following Jablonowski and Williamson (2006), l_2 error norms of the relativity vorticity field at model layer 23 are compared on the global grid as a function of time. Fig. 3 shows the l_2 norm for the SGL and MIX. In the initial stages of the model integration, the errors in the SGL simulations increased rapidly. By checking the original fields (figure not shown), it was found that numerous small-scale spurious fluctuations had emerged on both sides of the equator, the intensity of which was similar to the physically meaningful fluctuations.

After day 6, the primary fluctuations of the baroclinic waves in the SGL simulations began to develop, resembling the behaviour of the DBL simulations, and the errors started to decrease (Fig. 3). By day 10, the fluctuations rapidly developed, the primary fluctuations grew robustly, and the spurious fluctuations produced in the early stages of the SGL simulations also rapidly developed, leading to an increase in errors (Fig. 3). On day 11, the intensity of the spurious fluctuations developed in SGL was close to that of the primary fluctuations, which is unacceptable. Due to the slow growth of the

225 primary fluctuations in the early stages, the MIX simulation exhibited minimal errors before day 9 (Fig. 3). Subsequently, as the fluctuations matured rapidly, prominent differences in phase speed compared to the DBL emerged, leading to a rapid increase in errors.

4.2 Splitting supercell thunderstorms

The splitting supercell test of DCMIP2016 (Klemp et al. 2015; Zarzycki et al. 2019) emphasizes the importance of
230 scrutinizing non-hydrostatic model simulations of small-scale dynamics, especially as models approach spatial resolutions on the (sub) kilometre scale. This test utilized the small-planet testing framework (Wedi and Smolarkiewicz 2009), a cost-effective approach by scaling down Earth's radius by a factor of 120. The model employs the Kessler warm-rain microphysics scheme for simplified physics. This particular test case is characterized by unstable atmospheric conditions conducive to moist convection, posing a challenge numerical accuracy and stability. Klemp et al. (2015) suggested that an
235 increase in horizontal resolution should lead to convergent solutions. For GRIST, this behaviour has been verified by Zhang et al. (2020). Our investigation further examines the capability of the MIX configuration to accurately replicate the behaviours observed in the DBL simulations.

Figure 4 shows the q_r mixing ratio at 5 km elevation in both DBL and MIX simulations at four resolution choices (G4: ~4km, G6: ~1km, G7: ~0.5km and G8: ~0.25km). The DBL and MIX solutions show bulk similarities across all the
240 resolutions. At 7200s, a single updraft splits and evolves into a symmetric storm propagating towards the poles, with two supercells located ~30° from the equator. These supercells show subtle differences in their structure and intensity. At a low resolution of 4 km, the differences between MIX and DBL simulations are minimal at all altitudes (Figs. 4b-c). As the resolution increases from 4 km to 1km and to 0.5 km, the structural differences in supercells gradually become more pronounced (Figs. 4b-d, f-h, j-l). However, when the resolution further increases from 0.5 km to 0.25 km, the differences
245 diminish (Figs. 4n-p). For DBL simulation results, the differences between 0.5 km and 0.25 km are smaller than those between 1 km and 0.5 km, indicating that the solution converges almost at a resolution of 0.5km. At 0.25 km, the results of MIX simulation show greater similarity to those of DBL simulation at all altitudes (Figs. 4n-p). This indicates that, in mixed-precision simulation, supercells also achieve good convergence at this resolution, and thus the sensitivity to the precision level diminishes from 0.5 km to 0.25 km.

250 Figure 5 shows the maximum vertical speed and area-integrated rainfall rate over the global domain as a function of time for each resolution. The vertical speed in both MIX and DBL increases with resolution (Figs. 5a). From the start of the model integration until 5400s, the vertical speed curves of MIX and DBL simulations nearly overlap (Figs. 5a). After 5400s, a noticeable deviation appears, expect for the G4 grid. The difference in vertical speed between MIX and DBL is minimal at 4 km resolution, followed by the 0.25 km resolution, while it is larger at 1 km and 0.5 km resolutions (Figs. 5a). The area-
255 integrated rainfall rate curves exhibit similar evolutionary features (Figs. 5b). At very low resolutions, such as 4 km, the differences between MIX and DBL simulations are not significant. At a higher resolution of 0.25 km, the overall behaviour of supercells in MIX simulations is closer to that of DBL compared to 0.5km and 1 km resolutions. Both MIX and DBL solutions demonstrate good convergence.

4.3 Idealized tropical cyclone

This idealized tropical cyclone scenario integrates a three-dimensional dynamical core with a simple physics suite (Reed and Jablonowski 2012), alongside an analytic vortex initialization technique (Reed and Jablonowski 2011). The experiment produces the evolution of a tropical cyclone from a nascent, idealized vortex, highlighting the model's sensitivity to various parameter adjustments. Notably, alterations in tracer transport schemes in GRIST can produce subtle sensitivities in the development of the tropical cyclone due to the pressure gradient terms (Zhang et al. 2020), thereby establishing this case useful for assessing model precision sensitivity.

Figure 6 displays the wind speed at day 10 for the DBL (Figs. 6a and 6b) and MIX (Figs. 6c and 6d) simulations at G8 resolution. Fig. 6 (left) shows the longitude-height cross sections of the magnitude of the wind through the centre latitude of the vortex. Fig. 6 (right) displays the horizontal cross sections of the magnitude of the wind at surface model layer. The centre of vortex is defined as the grid point with the minimum surface pressure. At the day 10, the developed storm resembled a tropical cyclone. The overall behaviour in the MIX simulation was similar to that in the DBL simulation, with maximum winds near the surface and a distinct eyewall structure (Fig. 6). However, there was some differences in the vertical structure and centre location of the cyclone (Figs. 6a and 6c). In the MIX simulation, the generated cyclone was stronger, with higher wind speeds near the surface (Fig. 6c). The eyewall of the cyclone in the MIX simulation appeared less pronounced compared to that in the DBL simulation, where the cyclone's eyewall is narrower and straighter (Fig. 6c). Overall, the characteristics of the cyclone were comparable between the MIX and DBL simulations.

In addition to two deterministic control simulations using both double-precision and mixed-precision with the non-hydrostatic solver, eight ensemble simulations with the double-precision non-hydrostatic solver are further performed. This assesses the MIX simulation within the uncertainty range of the DBL simulation. The uncertainty range is quantified by the ensemble simulations encompassing eight initial-value perturbation members. Random small-amplitude perturbations were applied to the initial wind speeds (e.g., Li et al. 2020), where perturbations to the normal velocity at cell edges were prescribed within a range of 2% of their values in the control experiment.

Figure 7 describes the tracks of tropical cyclones, along with the evolution of minimum surface pressure and maximum surface wind speed over time. The red and blue lines represent two deterministic simulations conducted using MIX and DBL solvers, respectively. The eight random perturbation simulations with DBL solver are represented by gray lines. Minimal spread is observed in the early stages of the simulations (Figs. 7). Cyclone track separation between MIX and DBL simulations occurs at the day 1 (Fig. 7a). Subsequently, spread in the simulations increases over time (Figs. 7). The evolution of minimum surface pressure and maximum surface wind speed over time exhibits similar trends (Figs. 7b and 7c). No discernible difference is found between the sensitivity introduced by the MIX and that introduced by perturbed initial values in the DBL simulations. The overall behaviour of the MIX simulation falls within the range of uncertainty of the DBL simulation.

4.4 Atmospheric Model Intercomparison Project (AMIP) simulation

Following the establishment of the MIX dynamical core, a detailed examination of its integration with the model physics suite (Li et al. 2023) becomes crucial. The nonlinear interactions between the model's dynamics and its physical processes can result in varied performances across weather and climate simulations. It's imperative to investigate these differences to ensure that MIX simulations can accurately mirror the outcomes of DBL simulations in practical applications.

In assessing a new formulation for real-world modelling, our guiding principle is to first run long-term AMIP simulations (Zhang et al. 2021). This ensures that the model can achieve statistical equilibrium, maintain a realistic model climate, and has good integral properties such as conservation and balanced budgets (e.g., Fu et al. 2024). Subsequently, the same model, with minimal application-specific modifications, undergoes shorter-range but higher-resolution, kilometre-scale tests (Zhang et al. 2022).

The AMIP experiment is conducted in alignment with Zhang et al. (2021). This involved running both global hydrostatic and non-hydrostatic models with the weather physics suite at a G6 over a decade, spanning from 2001 to 2010. The simulations were performed under conditions with prescribed climatological sea surface temperatures and sea ice concentrations. The focus was narrowed to precipitation, which is a comprehensive metric due to its sensitivity to both model dynamics and physics, effectively reflecting the non-linear interactions that are crucial for accurate weather and climate simulations (Zhang and Chen 2016).

Figure 8 shows the simulated climatological (2001-2010) precipitation field for June-July-August (JJA) and December-January-February (DJF). Both the MIX hydrostatic and non-hydrostatic solvers are capable of replicating the JJA and DJF precipitation patterns in the DBL simulations. The discrepancies between MIX and DBL simulations are similar in both hydrostatic and non-hydrostatic simulations, with the primary differences occurring in the tropics. The precipitation differences shift from north to south along with the main rain bands as the season transition from summer to winter. The deviation in summer precipitation is greater than that in winter precipitation, because convective activities are most vigorous. In the summer, the MIX simulation overestimates the precipitation in the tropical coastal regions of the Western Pacific, especially along the western coast (Figs. 8a and 8b). In winter, the main biases in the MIX simulation are concentrated in the Southern Ocean (Figs. 8c and 8d).

These results may have two implications. In MIX simulations, the cumulative effects of rounding errors might be progressively magnified over the course of long-term climate integrations. This phenomenon could lead to notable differences in the simulated large-scale atmospheric phenomena. This contrasts with high-resolution shorter-range weather modeling, where discrepancies primarily emerge at the small scales, as will be discussed in Section 4.5. This might imply that MIX may more diverge from their DBL counterparts over extended integration periods, necessitating a careful consideration of how rounding errors accumulate and their impact on the climate simulation performance.

Second, the differences induced by varying levels of precision can be further exacerbated by physical processes within the climate system. A clear example is observed in the tropical regions during the boreal summer, where higher discrepancies are noted. This suggests that certain atmospheric conditions or regions, such as the tropics during periods of

intense solar heating, may be more susceptible to the effects of precision-level differences. These conditions can amplify the inherent precision differences, leading to more pronounced variations.

In the MIX implementation, Eq. 3 implies that global mass is conserved at the double precision level. The local mass flux is only conserved at the single precision level, because the mass flux and its divergence are treated as single precision. As mentioned in Section 2.3, we have retained a capability to compute the terms related to the mass flux divergence equation also in the double precision. Local mass can be conserved at the double precision level as well. We then evaluated the long-term climate integration results based on the hydrostatic solver.

Figure 9 shows the differences of the climatological precipitation field between MIX with single- (MIX_SGL_mass) and double-precision (MIX_DBL_mass) mass flux divergence against the pure DBL simulation. In the summer, the simulation differences between the MIX_SGL_mass and MIX_DBL_mass solver for the continuous equations is small (Figs. 9a and 9b). In the winter, the deviations in the MIX_SGL_mass is smaller than those in the MIX_DBL_mass solver (Figs. 9c and 9d). The deviations are most pronounced in the tropical convective precipitation over the Southern Ocean (Figs. 9c and 9d). The larger difference between MIX_DBL_mass and DBL is *likely* due to implicit type conversions, as discussed in Section 2.2.

4.5 A global storm-resolving simulation

Under the constraints of today's computational resources, executing global storm-resolving nonhydrostatic simulations remains resource intensive (Satoh et al. 2017; Stevens et al. 2019). The use of MIX simulations presents a cost-effective solution to this challenge. However, it has been reported, for instance by Nakano et al. (2018), that as the resolution of the model increases, the difference between MIX and DBL may increase, especially for the smaller-scale flow features. This observation prompts a closer investigation into the performance of nonhydrostatic models at high-resolution modelling.

A global storm-resolving experiment at 5 km (G9B3) is performed using the MIX nonhydrostatic model, following Zhang et al. (2022). The model was integrated from UTC00, 10th, to UTC00, 15th, July, 2015. We expect that the developed mixed-precision dynamical core can replicate the behaviour of DBL in kilometre-scale weather simulations.

Figure 10 show the period-accumulated precipitation (UTC00, 10th-UTC23, 15th, 2015) from the MIX and DBL model runs. All data have been interpolated to the $0.5°$ regular latitude-longitude grid. The precipitation pattern simulated by MIX are very close to those of DBL simulations. MIX obtains nearly the same general position, orientation, and intensification of the rain band (Figs. 10a, 10b). MIX and DBL also produced very comparable kinetic energy spectra (figure not shown).

Like the AMIP simulations, the differences in precipitation are primarily located within the tropics, with the most pronounced differences in areas with vigorous convection. Close-ups of these locations reveal that it is small-scale (a few grid spacing) that is most sensitive to the precision level, because small scales are most sensitive to numerical discretization and dissipation (Jablonowski and Williamson 2011). Considering that global meso-scale forecast at a few kilometres would greatly benefit from ensemble prediction (Palmer 2019), in practice, the MIX induced small-scale sensitivity may also fall within the uncertainty range of the ensemble, similar to that in Section 4.3.

5 Discussion

As mentioned in Section 2.3, the advective parts of the equations are not sensitive to the precision level and they are using higher-order operators. To understand whether this optimization outcome is sensitive to the nominal order of numerics, we utilized an isolated 3D tracer transport experiment (Kent et al. 2013, Hadley-like meridional circulation) and performed a convergence test. This test case was also performed by Zhang et al. (2020) and Li and Zhang (2022), and thus their results can be used as reference. By adjusting the very small values introduced by the limiter to be within the single-precision range, this equation (Eq. 10 in Fig.1) is solved independently in single-precision ("SPCODE" compiler option in the model code).

We used various orders of horizontal flux operators, namely RK3O2, RK3O3, and RK3O4 (combinations of third order Runge-Kutta integration scheme and nominal 2nd- to 4th-order spatial flux operators). RK3O3 is used in other tests of this paper. The vertical advection operator remains unchanged. The results are shown in Fig. 11. The tested resolutions and associated time step include G5L30 (600s), G6L60 (300s), G7L120 (150s), G8L240 (75s) and G9L480 (37.5s). The results demonstrate that, using different orders of horizontal flux operators, the single-precision simulations are comparable to the double-precision simulations across all resolutions, with nearly identical error norms and convergence rates.

This outcome suggests that, within the current code implementation, the advective part of the model demonstrates greater resilience when subjected to changes in precision, regardless of the nominal order of the operator. This supports the optimization results in the entire dynamical equations. As reported by Nakano et al. (2018) and Yin et al. (2020), the precision-sensitive components are related to the specific numerical algorithms. Meanwhile, badly conditioned code or poor coding practice can also necessitate double-precision calculations (Váňa et al. 2016; Palmer 2020). For other components (e.g., pressure gradient) that currently have higher sensitivity to the precision level, we believe that it may require more careful code implementation to allow us to benefit more from reduced-precision computing.

6 Summary and outlook

In this study, we investigated mixed-precision computing within the GRIST dynamical core, identifying the equation terms particularly sensitive to numerical precision. We outlined an optimization procedure characterized by a limited extent of iterative development. Given the current development trajectory of high-performance computing, where advancements in memory bandwidth lag behind peak processor performance improvements, mixed-precision computation holds promise for enhancing weather and climate model development. The major conclusions are summarized as follows.

We discovered that terms sensitive to numerical precision primarily involve pressure gradient and gravity. In contrast, advective terms exhibit resilience to single precision and can be optimized. The advective terms are computationally more expensive than the pressure gradient and gravity terms. The viability of employing mixed-precision computing in the GRIST dynamical core has been validated across a spectrum of scenarios, from idealized flow to real-world AMIP and global storm-resolving simulations. These MIX experiments yielded results remarkably similar to those from DBL simulations. For dycore, the runtime for hydrostatic and non-hydrostatic solvers was reduced by 24% and 27%, respectively. The tracer

390   transport solver witnessed a runtime reduction of 44%. The overall time savings depend on the proportion of dycore and tracer transport in the total computation time, varying by application.

We noted a higher sensitivity to precision in long-term climate simulations compared to short-term higher-resolution weather forecasts, particularly affecting the precipitation field over certain regions. In weather forecast, the difference between MIX and DBL are mainly for the small scales, while in AMIP simulations, the difference is found for the larger

395   scales. These effects may primarily stem from the model sensitivity to the precision level or from biases introduced by mixed-precision computations themselves.

It also needs to recognize that while the mixed-precision GRIST dynamical core has been examined across multiple scenarios, more realistic tests and some fine tuning may still be needed to ensure robust operational forecasting. Some alternative advection schemes in the tracer transport module have not been implemented to single precision yet and this can

400   be achieved in future. It is also possible that room remains for further optimization and time reduction, especially when considering modifying the original code implementation and using further reduced significant digits.

Acknowledgments

Editors and reviewers are thanked for their handling and comments of this paper. This study is supported by a National Talent Project (2021).

405   **Code and Data availability**

Model code and plotting data related to this manuscript is available at: https://zenodo.org/records/11229770.

Author contribution

SYC developed the mixed precision model code and prepared the initial draft. YZ designed and led this model development research. YW contributed to experiments. All the authors discussed this work and contributed to the final manuscript version.

410   **Competing interests**

None.

Appendix

We define the three-dimensional global integral of $\mathcal{H}$ as:

$$I(\mathcal{H}) = \int_{z=z_{surface}}^{z=z_{top}} \oiint \mathcal{H} \, dAdz, \quad (A1)$$

415   where A denotes cell area and z denotes height. The vertical integral is omitted if two-dimensional space is under consideration. The definitions of L_1, L_2, and L_∞ are as follows:

$$L_1 = \frac{I(|\mathcal{H}-\mathcal{H}_T|)}{I(|\mathcal{H}_T|)}, \quad (A2)$$

$$L_2 = \sqrt{\frac{I[(\mathcal{H}-\mathcal{H}_T)^2]}{I[(\mathcal{H}_T)^2]}}, \quad (A3)$$

$$L_\infty = \frac{max\forall|\mathcal{H}-\mathcal{H}_T|}{max\forall|\mathcal{H}_T|}, \qquad (A4)$$

420   where $\mathcal{H}$ and $\mathcal{H}_T$ are the computational solution and true solution, respectively. Max $\forall$ means selecting the maximum value from the field.

References

Abdelfattah, A., Anzt, H., Boman, E. G., Carson, E., Cojean, T., Dongarra, J., Fox, A., Gates, M., Higham, N. J., Li, X. S., Loe, J., Luszczek, P., Pranesh, S., Rajamanickam, S., Ribizel, T., Smith, B. F., Swirydowicz, K., Thomas, S.,
425   Tomov, S., Tsai, Y. M., and Yang, U. M.: A survey of numerical linear algebra methods utilizing mixed-precision arithmetic, The International Journal of High Performance Computing Applications, 35, 344-369, doi:10.1177/10943420211003313, 2021.

Baboulin, M., Buttari, A., Dongarra, J., Kurzak, J., Langou, J., Langou, J., Luszczek, P., and Tomov, S.: Accelerating scientific computations with mixed precision algorithms, Computer Physics Communications, 180, 2526-2533,
430   doi:https://doi.org/10.1016/j.cpc.2008.11.005, 2009.

Bauer, P., Dueben, P. D., Hoefler, T., Quintino, T., Schulthess, T. C., and Wedi, N. P.: The digital revolution of Earth-system science, Nature Computational Science, 1, 104-113, doi:10.1038/s43588-021-00023-0, 2021.

Benjamin, S. G., Brown, J. M., Brunet, G., Lynch, P., Saito, K., and Schlatter, T. W.: 100 Years of Progress in Forecasting and NWP Applications, Meteorological Monographs, 59, 13.11-13.67,
435   doi:10.1175/AMSMONOGRAPHS-D-18-0020.1, 2019.

Brogi, F., Bnà, S., Boga, G., Amati, G., Esposti Ongaro, T., and Cerminara, M.: On floating point precision in computational fluid dynamics using OpenFOAM, Future Generation Computer Systems, 152, 1-16, doi:https://doi.org/10.1016/j.future.2023.10.006, 2024.

Chantry, M., Thornes, T., Palmer, T., and Düben, P.: Scale-Selective Precision for Weather and Climate Forecasting,
440   Monthly Weather Review, 147, 645-655, doi:https://doi.org/10.1175/MWR-D-18-0308.1, 2019.

Cotronei, A. and Slawig, T.: Single-precision arithmetic in ECHAM radiation reduces runtime and energy consumption, Geosci. Model Dev., 13, 2783-2804, doi:10.5194/gmd-13-2783-2020, 2020.

Düben, P. D., McNamara, H., and Palmer, T. N.: The use of imprecise processing to improve accuracy in weather & climate prediction, Journal of Computational Physics, 271, 2-18, doi:https://doi.org/10.1016/j.jcp.2013.10.042,
445   2014.

Düben, P. D. and Palmer, T. N.: Benchmark Tests for Numerical Weather Forecasts on Inexact Hardware, Monthly Weather Review, 142, 3809-3829, doi:https://doi.org/10.1175/MWR-D-14-00110.1, 2014.

Düben, P. D., Russell, F. P., Niu, X., Luk, W., and Palmer, T. N.: On the use of programmable hardware and reduced numerical precision in earth-system modeling, Journal of Advances in Modeling Earth Systems, 7, 1393-1408, doi:10.1002/2015MS000494, 2015.

Fornaciari, W., Agosta, G., Cattaneo, D., Denisov, L., Galimberti, A., Magnani, G., and Zoni, D.: Hardware and Software Support for Mixed Precision Computing: a Roadmap for Embedded and HPC Systems, 2023 Design, Automation & Test in Europe Conference & Exhibition (DATE), 2023, 1-6,

Fu, H., Liao, J., Ding, N., Duan, X., Gan, L., Liang, Y., Wang, X., Yang, J., Zheng, Y., Liu, W., Wang, L., and Yang, G.: Redesigning CAM-SE for peta-scale climate modeling performance and ultra-high resolution on Sunway TaihuLight, Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, Denver, Colorado, 2017.

Fu, Z., Zhang, Y., Li, X., and Rong, X.: Intercomparison of Two Model Climates Simulated by a Unified Weather-Climate Model System (GRIST), Part I: Mean State, Climate Dynamics, 2024.

Gan, L., Fu, H., Luk, W., Yang, C., Xue, W., Huang, X., Zhang, Y., and Yang, G.: Accelerating solvers for global atmospheric equations through mixed-precision data flow engine, 2013 23rd International Conference on Field programmable Logic and Applications, 2013, 1-6,

Gu, J., Feng, J., Hao, X., Fang, T., Zhao, C., An, H., Chen, J., Xu, M., Li, J., Han, W., Yang, C., Li, F., and Chen, D.: Establishing a non-hydrostatic global atmospheric modeling system at 3-km horizontal resolution with aerosol feedbacks on the Sunway supercomputer of China, Science Bulletin, doi:https://doi.org/10.1016/j.scib.2022.03.009, 2022.

Harris, L. M. and Lin, S.-J.: A Two-Way Nested Global-Regional Dynamical Core on the Cubed-Sphere Grid, Monthly Weather Review, 141, 283-306, doi:10.1175/MWR-D-11-00201.1, 2012.

Jablonowski, C. and Williamson, D. L.: A baroclinic instability test case for atmospheric model dynamical cores, Quarterly Journal of the Royal Meteorological Society, 132, 2943-2975, doi:10.1256/qj.06.12, 2006.

Jablonowski, C. and Williamson, D.: The Pros and Cons of Diffusion, Filters and Fixers in Atmospheric General Circulation Models, In: Lauritzen, P. H., C. Jablonowski, M. A. Taylor, R. D. Nair (Eds.), Numerical Techniques for Global Atmospheric Models, Lecture Notes in Computational Science and Engineering, Springer, Vol. 80, 381-493, 2011.

Kent, J., Ullrich, P. A., and Jablonowski, C.: Dynamical core model intercomparison project: Tracer transport test cases, Quarterly Journal of the Royal Meteorological Society, 140, 1279-1293, doi:10.1002/qj.2208, 2013.

Klemp, J. B., Skamarock, W. C., and Park, S. H.: Idealized global nonhydrostatic atmospheric test cases on a reduced-radius sphere, Journal of Advances in Modeling Earth Systems, 7, 1155-1177, doi:10.1002/2015MS000435, 2015.

Li, J. and Zhang, Y.: Enhancing the stability of a global model by using an adaptively implicit vertical moist transport scheme, Meteorology and Atmospheric Physics, 134, 55, doi:10.1007/s00703-022-00895-5, 2022.

Li, X., Peng, X., and Zhang, Y.: Investigation of the effect of the time step on the physics–dynamics interaction in CAM5 using an idealized tropical cyclone experiment, Climate Dynamics, 55, 665-680, doi:10.1007/s00382-020-05284-5, 2020.

Li, X., Zhang, Y., Peng, X., Zhou, B., Li, J., and Wang, Y.: Intercomparison of the weather and climate physics suites of a unified forecast–climate model system (GRIST-A22.7.28) based on single-column modeling, Geosci. Model Dev., 16, 2975-2993, doi:10.5194/gmd-16-2975-2023, 2023.

Maynard, C. M. and Walters, D. N.: Mixed-precision arithmetic in the ENDGame dynamical core of the Unified Model, a numerical weather prediction and climate model code, Computer Physics Communications, 244, 69-75, doi:https://doi.org/10.1016/j.cpc.2019.07.002, 2019.

Nakano, M., Yashiro, H., Kodama, C., and Tomita, H.: Single Precision in the Dynamical Core of a Nonhydrostatic Global Atmospheric Model: Evaluation Using a Baroclinic Wave Test Case, Monthly Weather Review, 146, 409-416, doi:10.1175/MWR-D-17-0257.1, 2018.

Palmer, T.: The ECMWF ensemble prediction system: Looking back (more than) 25 years and projecting forward 25 years, Quarterly Journal of the Royal Meteorological Society, 145, 12-24, doi:https://doi.org/10.1002/qj.3383, 2019.

Palmer, T. N.: The physics of numerical analysis: a climate modelling case study, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 378, 20190058, doi:doi:10.1098/rsta.2019.0058, 2020.

Reed, K. A. and Jablonowski, C.: An Analytic Vortex Initialization Technique for Idealized Tropical Cyclone Studies in AGCMs, Monthly Weather Review, 139, 689-710, doi:10.1175/2010mwr3488.1, 2011.

Reed, K. A. and Jablonowski, C.: Idealized tropical cyclone simulations of intermediate complexity: A test case for AGCMs, Journal of Advances in Modeling Earth Systems, 4, doi:10.1029/2011MS000099, 2012.

Santos, F. F. D., Carro, L., Vella, F., and Rech, P.: Assessing the Impact of Compiler Optimizations on GPUs Reliability, ACM Trans. Archit. Code Optim., 21, Article 26, doi:10.1145/3638249, 2024.

Satoh, M., Tomita, H., Yashiro, H., Kajikawa, Y., Miyamoto, Y., Yamaura, T., Miyakawa, T., Nakano, M., Kodama, C., Noda, A. T., Nasuno, T., Yamada, Y., and Fukutomi, Y.: Outcomes and challenges of global high-resolution non-hydrostatic atmospheric simulations using the K computer, Progress in Earth and Planetary Science, 4, 13, doi:10.1186/s40645-017-0127-8, 2017.

Sergeev, D. E., Mayne, N. J., Bendall, T., Boutle, I. A., Brown, A., Kavčič, I., Kent, J., Kohary, K., Manners, J., Melvin, T., Olivier, E., Ragta, L. K., Shipway, B., Wakelin, J., Wood, N., and Zerroukat, M.: Simulations of idealised 3D atmospheric flows on terrestrial planets using LFRic-Atmosphere, Geosci. Model Dev., 16, 5601-5626, doi:10.5194/gmd-16-5601-2023, 2023.

Skamarock, W. C., Klemp, J. B., Duda, M. G., Fowler, L. D., Park, S.-H., and Ringler, T. D.: A Multiscale Nonhydrostatic Atmospheric Model Using Centroidal Voronoi Tesselations and C-Grid Staggering, Monthly Weather Review, 140, 3090-3105, doi:10.1175/MWR-D-11-00215.1, 2012.

Stevens, B., Satoh, M., Auger, L., Biercamp, J., Bretherton, C. S., Chen, X., Düben, P., Judt, F., Khairoutdinov, M., Klocke, D., Kodama, C., Kornblueh, L., Lin, S.-J., Neumann, P., Putman, W. M., Röber, N., Shibuya, R., Vanniere, B., Vidale, P. L., Wedi, N., and Zhou, L.: DYAMOND: the DYnamics of the Atmospheric general circulation Modeled On Non-hydrostatic Domains, Progress in Earth and Planetary Science, 6, 61, doi:10.1186/s40645-019-0304-z, 2019.

Taylor, M., Caldwell, P. M., Bertagna, L., Clevenger, C., Donahue, A., Foucar, J., Guba, O., Hillman, B., Keen, N., Krishna, J., Norman, M., Sreepathi, S., Terai, C., White, J. B., Salinger, A. G., McCoy, R. B., Leung, L.-y. R., Bader, D. C., and Wu, D.: The Simple Cloud-Resolving E3SM Atmosphere Model Running on the Frontier Exascale System, Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, <conf-loc>, <city>Denver</city>, <state>CO</state>, <country>USA</country>, </conf-loc>, 2023.

Thornes, T., Düben, P., and Palmer, T.: On the use of scale-dependent precision in Earth System modelling, Quarterly Journal of the Royal Meteorological Society, 143, 897-908, doi:https://doi.org/10.1002/qj.2974, 2017.

Thuburn, J.: Some conservation issues for the dynamical cores of NWP and climate models, Journal of Computational Physics, 227, 3715-3730, doi:http://dx.doi.org/10.1016/j.jcp.2006.08.016, 2008.

Tomita, H. and Satoh, M.: A new dynamical framework of nonhydrostatic global model using the icosahedral grid, Fluid Dynamics Research, 34, 357, 2004.

Ullrich, P. A., Melvin, T., Jablonowski, C., and Staniforth, A.: A proposed baroclinic wave test case for deep- and shallow-atmosphere dynamical cores, Quarterly Journal of the Royal Meteorological Society, 140, 1590-1602, doi:10.1002/qj.2241, 2014.

Váňa, F., Düben, P., Lang, S., Palmer, T., Leutbecher, M., Salmond, D., and Carver, G.: Single Precision in Weather Forecasting Models: An Evaluation with the IFS, Monthly Weather Review, 145, 495-502, doi:10.1175/MWR-D-16-0228.1, 2016.

Wedi, N. P. and Smolarkiewicz, P. K.: A framework for testing global non-hydrostatic models, Quarterly Journal of the Royal Meteorological Society, 135, 469-484, doi:10.1002/qj.377, 2009.

Wedi, N. P., Polichtchouk, I., Dueben, P., Anantharaj, V. G., Bauer, P., Boussetta, S., Browne, P., Deconinck, W., Gaudin, W., Hadade, I., Hatfield, S., Iffrig, O., Lopez, P., Maciel, P., Mueller, A., Saarinen, S., Sandu, I., Quintino, T., and Vitart, F.: A Baseline for Global Weather and Climate Simulations at 1 km Resolution, Journal of Advances in Modeling Earth Systems, 12, e2020MS002192, doi:https://doi.org/10.1029/2020MS002192, 2020.

545   Yang, C., Xue, W., Fu, H., You, H., Wang, X., Ao, Y., Liu, F., Gan, L., Xu, P., Wang, L., Yang, G., and Zheng, W.:
10M-Core Scalable Fully-Implicit Solver for Nonhydrostatic Atmospheric Dynamics, SC '16: Proceedings of the
International Conference for High Performance Computing, Networking, Storage and Analysis, 2016, 57-68,

Yin, F., Song, J., Wu, J., and Zhang, W.: An implementation of single-precision fast spherical harmonic transform in
Yin–He global spectral model, Quarterly Journal of the Royal Meteorological Society, 147, 2323-2334,
550   doi:https://doi.org/10.1002/qj.4026, 2021.

Yu, R., Zhang, Y., Wang, J., Li, J., Chen, H., Gong, J., and Chen, J.: Recent Progress in Numerical Atmospheric
Modeling in China, Advances in Atmospheric Sciences, 36, 938-960, doi:10.1007/s00376-019-8203-1, 2019.

Zängl, G., Reinert, D., Rípodas, P., and Baldauf, M.: The ICON (ICOsahedral Non-hydrostatic) modelling framework of
DWD and MPI-M: Description of the non-hydrostatic dynamical core, Quarterly Journal of the Royal
555   Meteorological Society, 141, 563-579, doi:10.1002/qj.2378, 2015.

Zarzycki, C. M., Jablonowski, C., Kent, J., Lauritzen, P. H., Nair, R., Reed, K. A., Ullrich, P. A., Hall, D. M., Taylor, M.
A., Dazlich, D., Heikes, R., Konor, C., Randall, D., Chen, X., Harris, L., Giorgetta, M., Reinert, D., Kühnlein, C.,
Walko, R., Lee, V., Qaddouri, A., Tanguay, M., Miura, H., Ohno, T., Yoshida, R., Park, S. H., Klemp, J. B., and
Skamarock, W. C.: DCMIP2016: the splitting supercell test case, Geosci. Model Dev., 12, 879-892,
560   doi:10.5194/gmd-12-879-2019, 2019.

Zhang, Y. and Chen, H.: Comparing CAM5 and Superparameterized CAM5 Simulations of Summer Precipitation
Characteristics over Continental East Asia: Mean State, Frequency?Intensity Relationship, Diurnal Cycle, and
Influencing Factors, Journal of Climate, 29, 1067-1089, doi:10.1175/JCLI-D-15-0342.1, 2016.

Zhang, Y.: Extending High-Order Flux Operators on Spherical Icosahedral Grids and Their Applications in the
565   Framework of a Shallow Water Model, Journal of Advances in Modeling Earth Systems, 10, 145-164,
doi:10.1002/2017MS001088, 2018.

Zhang, Y., Li, J., Yu, R., Zhang, S., Liu, Z., Huang, J., and Zhou, Y.: A Layer-Averaged Nonhydrostatic Dynamical
Framework on an Unstructured Mesh for Global and Regional Atmospheric Modeling: Model Description, Baseline
Evaluation, and Sensitivity Exploration, Journal of Advances in Modeling Earth Systems, 11, 1685-1714,
570   doi:10.1029/2018MS001539, 2019.

Zhang, Y., Li, J., Yu, R., Liu, Z., Zhou, Y., Li, X., and Huang, X.: A Multiscale Dynamical Model in a Dry-Mass
Coordinate for Weather and Climate Modeling: Moist Dynamics and Its Coupling to Physics, Monthly Weather
Review, 148, 2671-2699, doi:10.1175/MWR-D-19-0305.1, 2020.

Zhang, Y., Yu, R., Li, J., Li, X., Rong, X., Peng, X., and Zhou, Y.: AMIP Simulations of a Global Model for Unified
575   Weather-Climate Forecast: Understanding Precipitation Characteristics and Sensitivity Over East Asia, Journal of
Advances in Modeling Earth Systems, 13, e2021MS002592, doi:https://doi.org/10.1029/2021MS002592, 2021.

Zhang, Y., Li, X., Liu, Z., Rong, X., Li, J., Zhou, Y., and Chen, S.: Resolution Sensitivity of the GRIST Nonhydrostatic Model From 120 to 5 km (3.75 km) During the DYAMOND Winter, Earth and Space Science, 9, e2022EA002401, doi:https://doi.org/10.1029/2022EA002401, 2022.

580   Zhang, Y., Li, J., Zhang, H., Li, X., Dong, L., Rong, X., Zhao, C., Peng, X., and Wang, Y.: History and Status of Atmospheric Dynamical Core Model Development in China, in: Numerical Weather Prediction: East Asian Perspectives, edited by: Park, S. K., Springer International Publishing, Cham, 3-36, 2023.

Table 1. Grid Name and Corresponding Horizontal Resolutions.

Grid level of a subdivided icosahedron	Horizontal resolution (km) on a full-size Earth and on a small-radius Sphere (if used)	Number of Cells
G4	480km (4km)	2,562
G5	240km	10,242
G6	120km (1km)	40,962
G7	60km (0.5km)	163,842
G8	30km (0.25km)	655,362
G9	15km	2,621,442
G9B3	5km	23,592,962

Table 2. Elapsed time using single-, mixed- and double precision **(**The runtime of each solver is normalized to that of the corresponding solver in double-precision**).**

Grid Name	Precision	Dycore time (1440 steps)	Tracer time (1440 steps)
G8	DBL	1 (NDC) 1 (HDC)	1
	SGL	0.53 (NDC) 0.56 (HDC)	0.58
	MIX	0.73 (NDC) 0.76 (HDC)	0.56

590

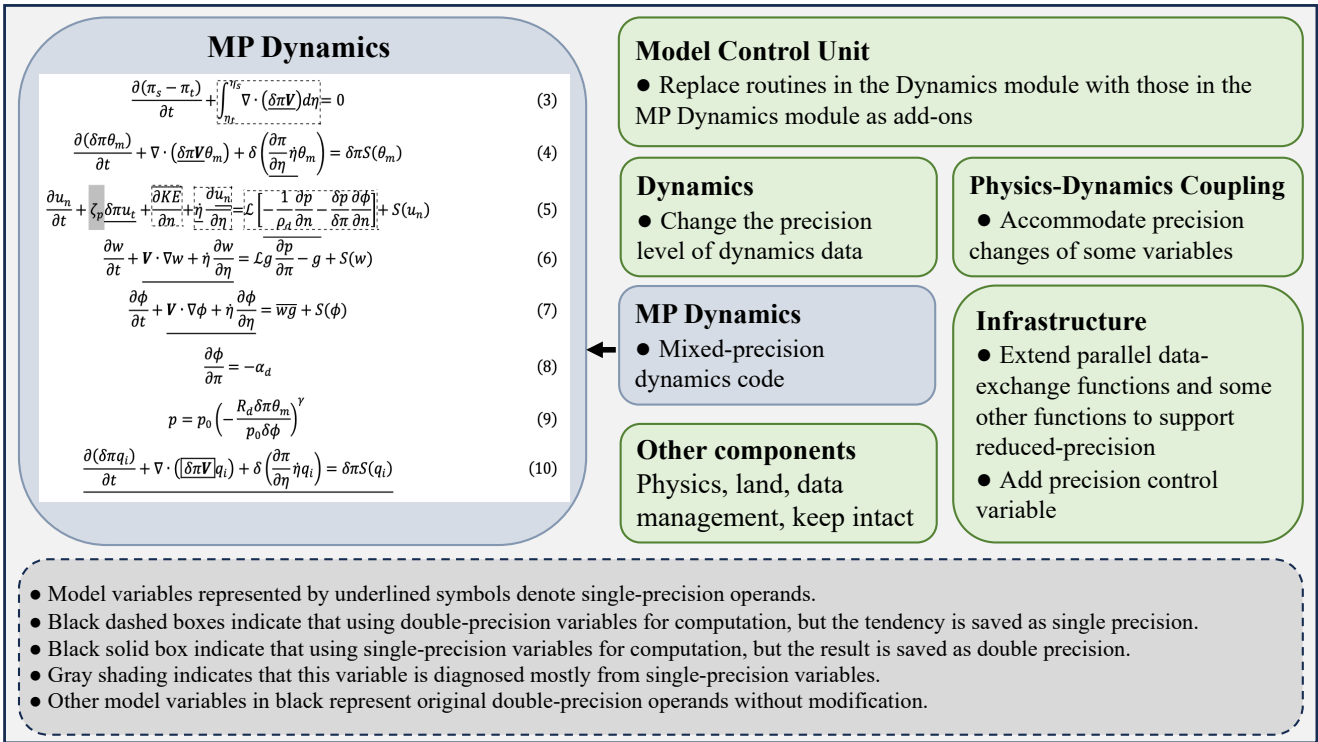



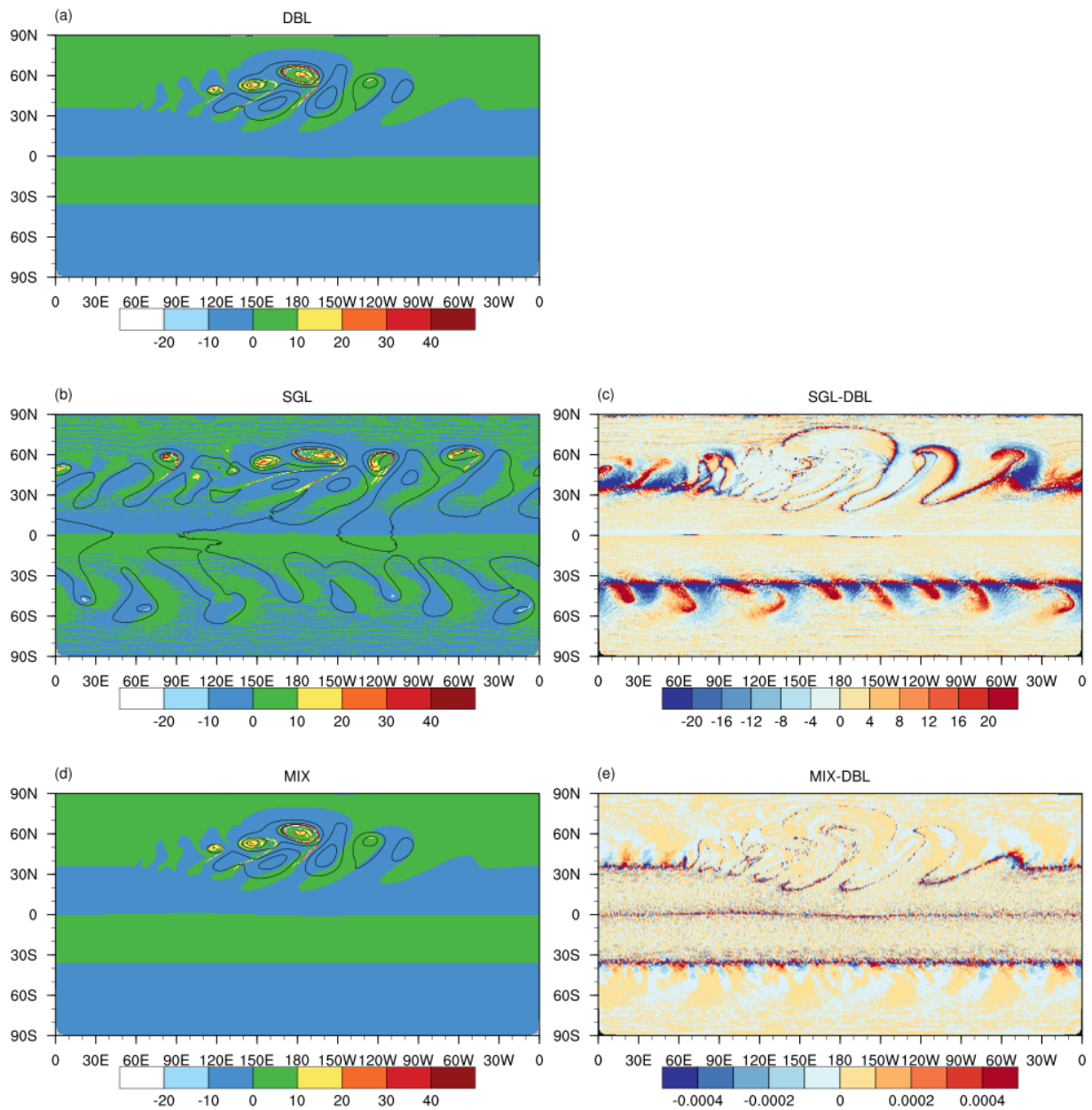
Figure1: Modifications to the GRIST model code repository for implementing the mixed-precision dynamical core.

595

Figure 2: Baroclinic wave development at day 11 in the (a) DBL simulation, (b) SGL simulation and (d) MIX simulation. (left) Colors show relative vorticity ($\times 10^{-5}\ s^{-1}$) and contours of the surface pressure and (right) the relative error between SGL and DBL, as well as the difference between MIX and DBL.

600

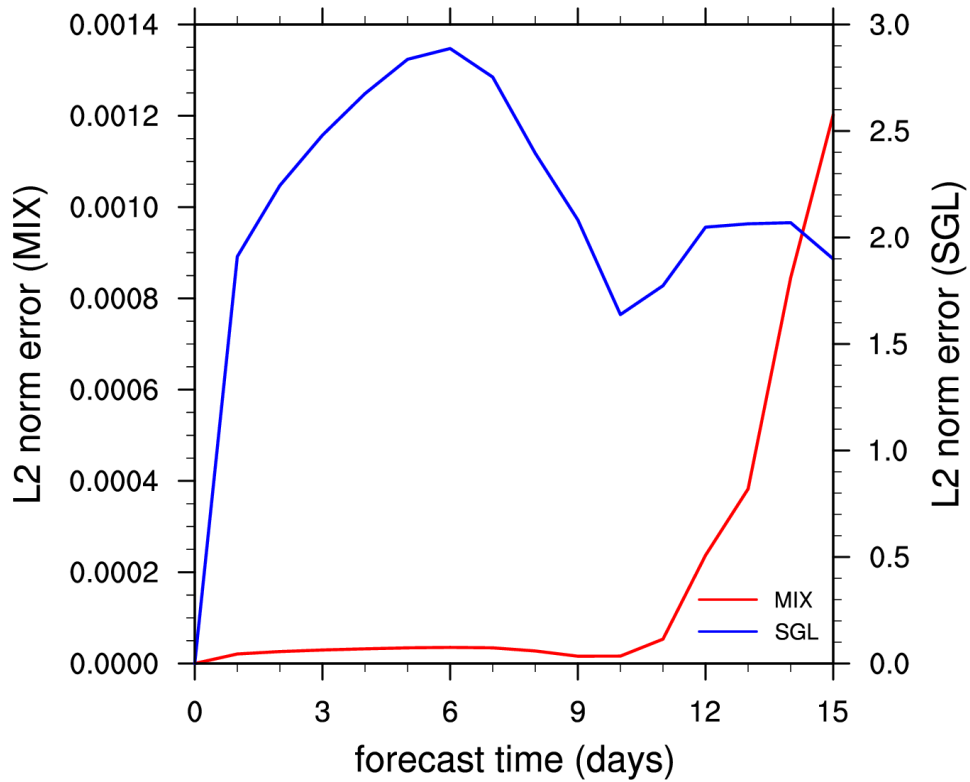



Figure 3: Time evolution of global l_2 difference norm of simulated relative vorticity between the SGL and DBL, as well as l_2 difference norm between the MIX and DBL. Red and blue represent SGL and MIX experiments, respectively.

605

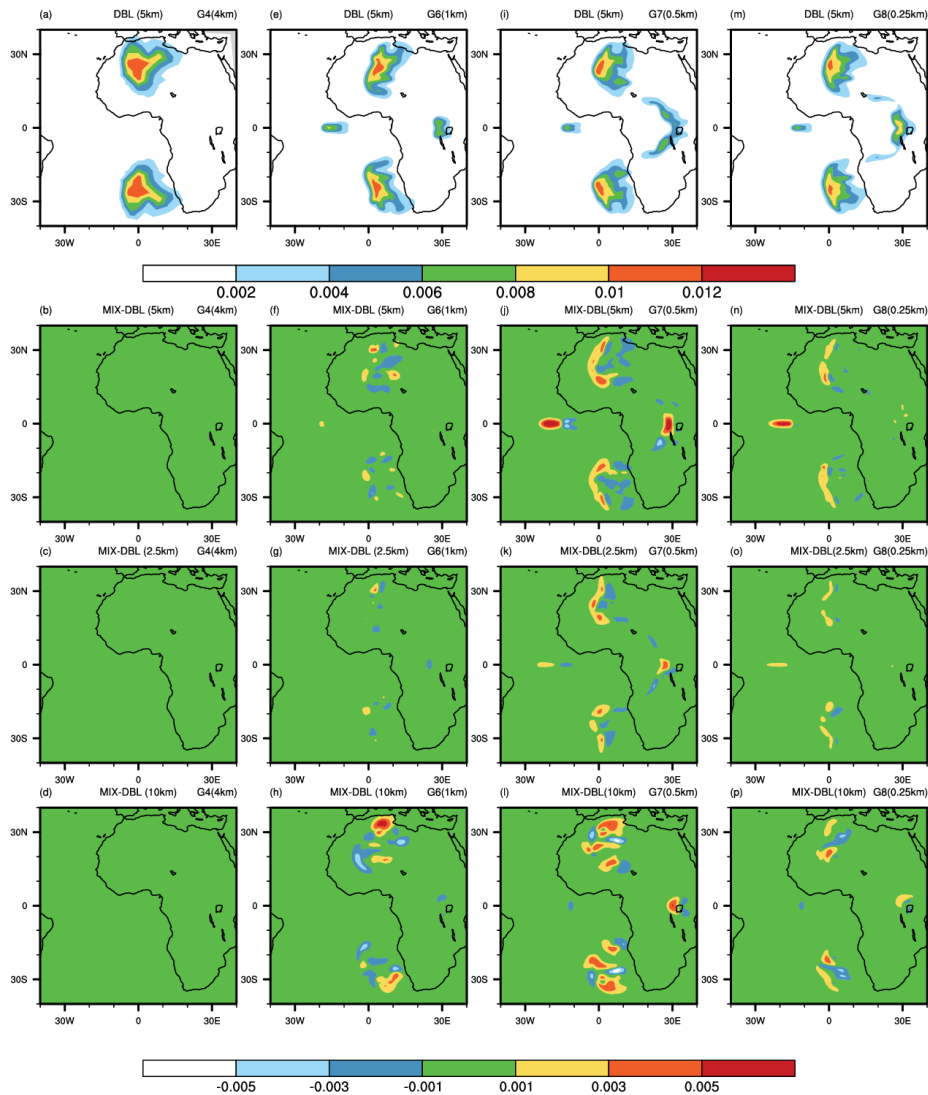



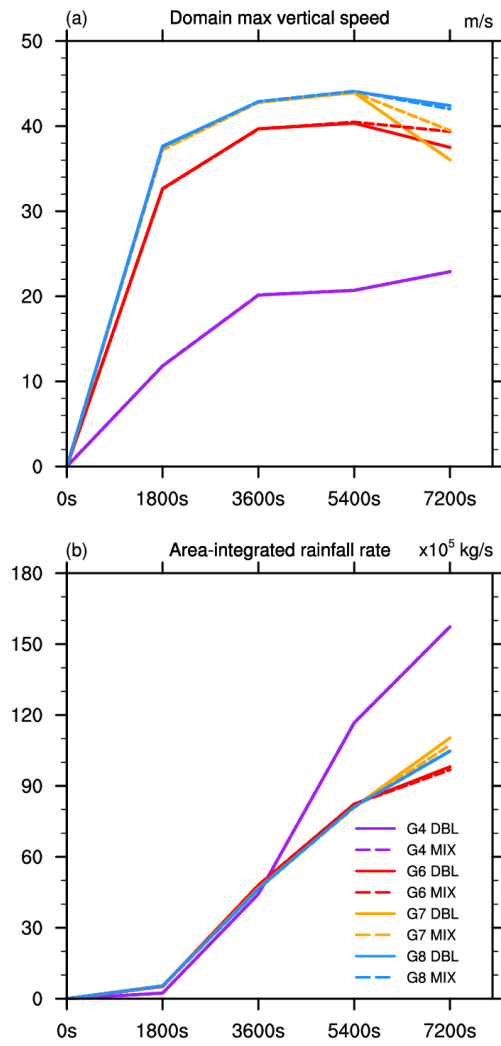
Figure 4: Horizontal cross-sections of rainwater mixing ratio at different heights from supercell thunderstorms simulations. The first row displays double-precision simulations at 5 km altitude. The second, third, and fourth rows show the differences between mixed-precision and double-precision simulations at 5 km, 2.5 km, and 10 km altitudes, respectively. The four columns represent results at different resolutions: G4 (4 km), G6 (1 km), G7 (0.5 km), and G8 (0.25 km) from left to right.

615

Figure 5: The (a) domain maximum vertical speed and (b) area-integrated rainfall rate obtained from the supercell simulations.

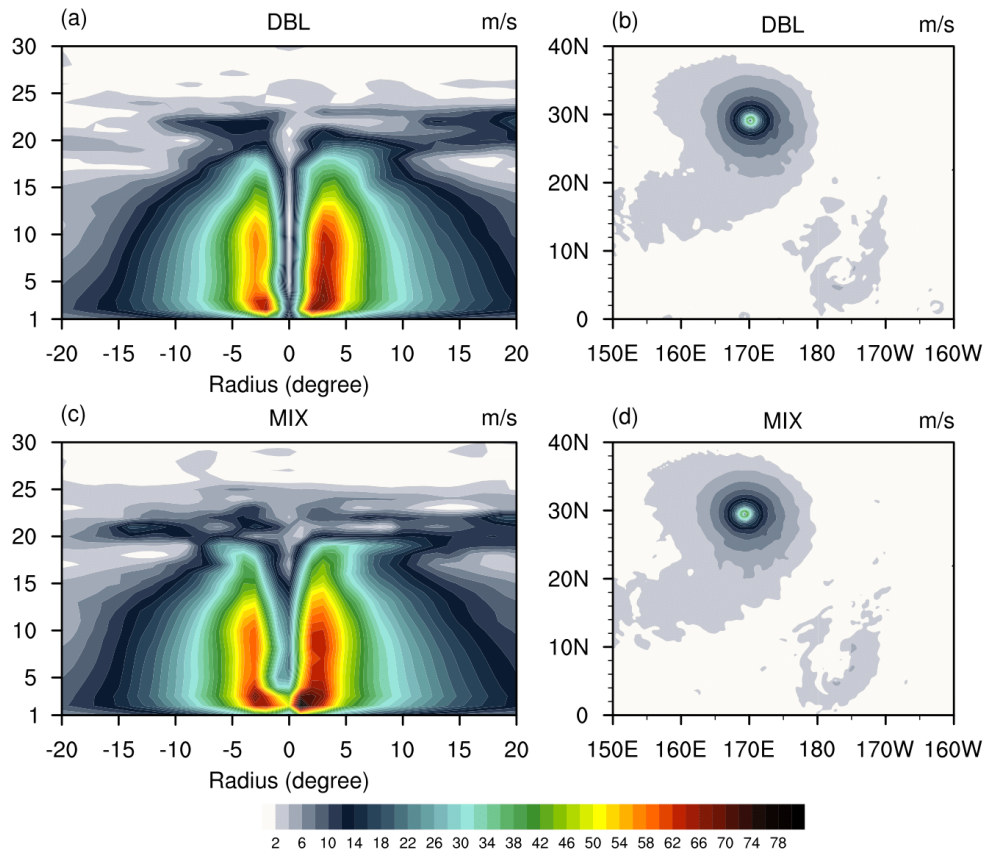



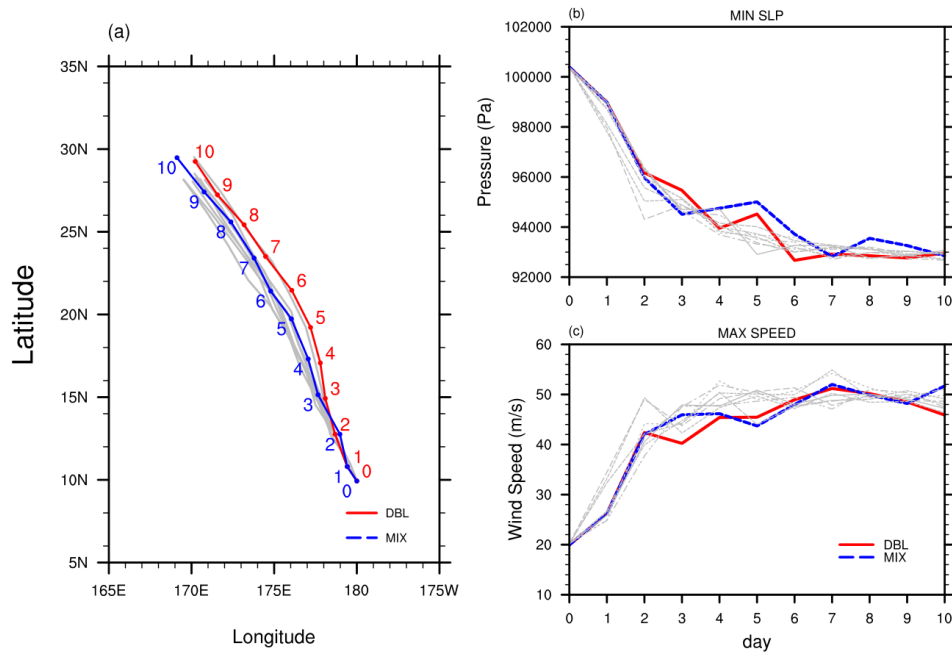
Figure 6: The simulated wind speed (m s⁻¹) at the G8 resolution with NDC solver, including MIX (a, b) and DBL (c, d) simulations. (left) Longitude-height cross section of the wind speed through the center latitude of the vortex as a function of the radius from the vortex center. (right) Horizontal cross section of the wind speed at surface model layer.

625

Figure 7: The results of deterministic and ensemble simulations. (a) The track of the tropical cyclone center for MIX (blue lines) and DBL (read lines) deterministic simulations. Time evolution of the (b) minimum surface pressure and (c) maximum surface wind speed from the and deterministic and ensemble simulations. The red and blue lines represent the deterministic MIX
630 **and DBL simulations, respectively. The gray lines represent the eight runs with random perturbations to initial normal velocity at cell edges.**

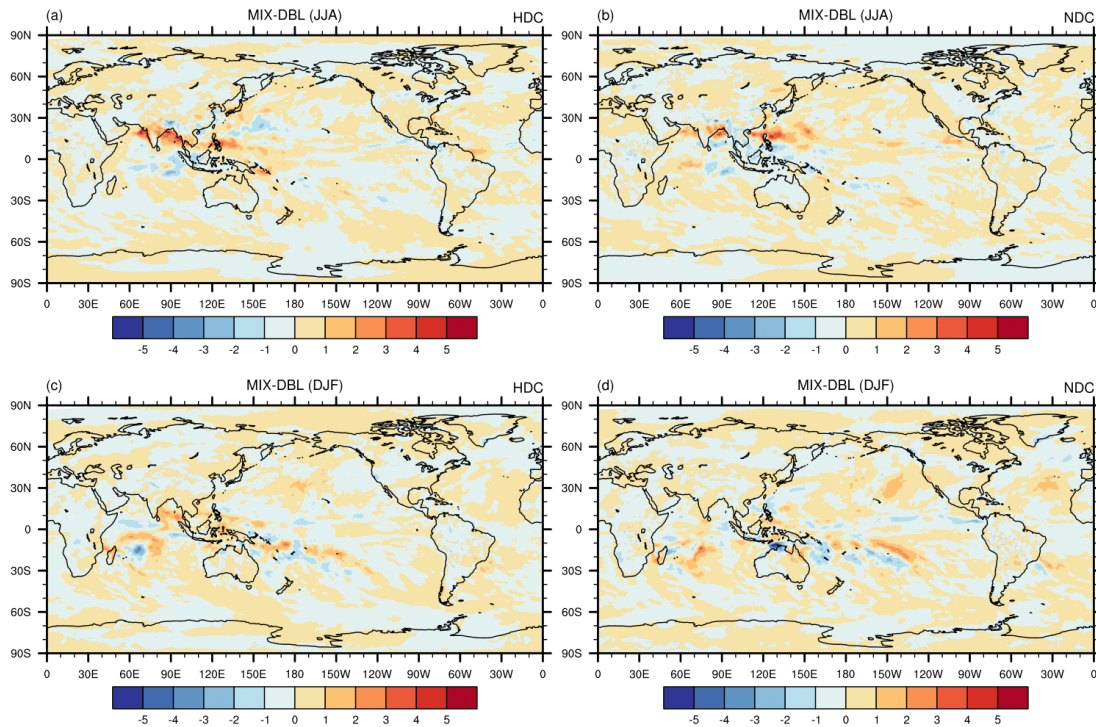



Figure 8: The difference between the MIX and DBL simulations, including solutions from the hydrostatic (left column) and non-hydrostatic (right column) solver. The first and second rows respectively display the averaged (2001-2010) precipitation rate (mm/day) for JJA and DJF.

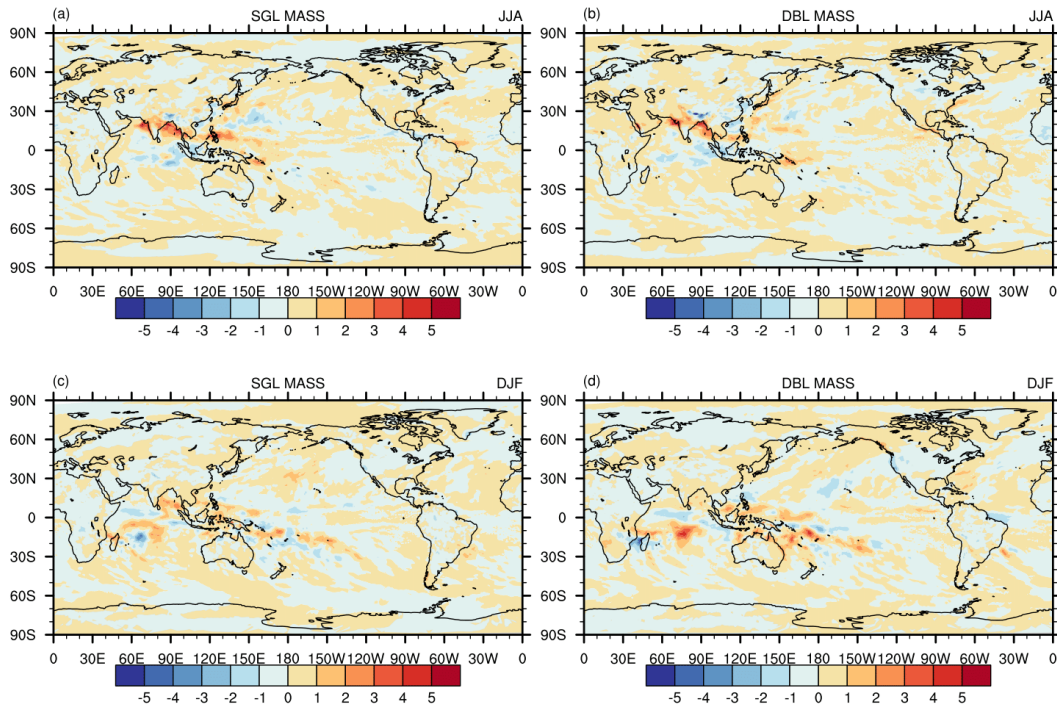



Figure 9: (a) Difference between the JJA-averaged (2001-2010) precipitation rate (mm/day) simulated by the SGL continuous equation solver in mixed-precision mode and the "true DBL value"; (b) same as (a) but for the DBL continuous equation solver. (c)-(d) same as (a)-(b) but for the DJF-averaged (2001-2010) results.

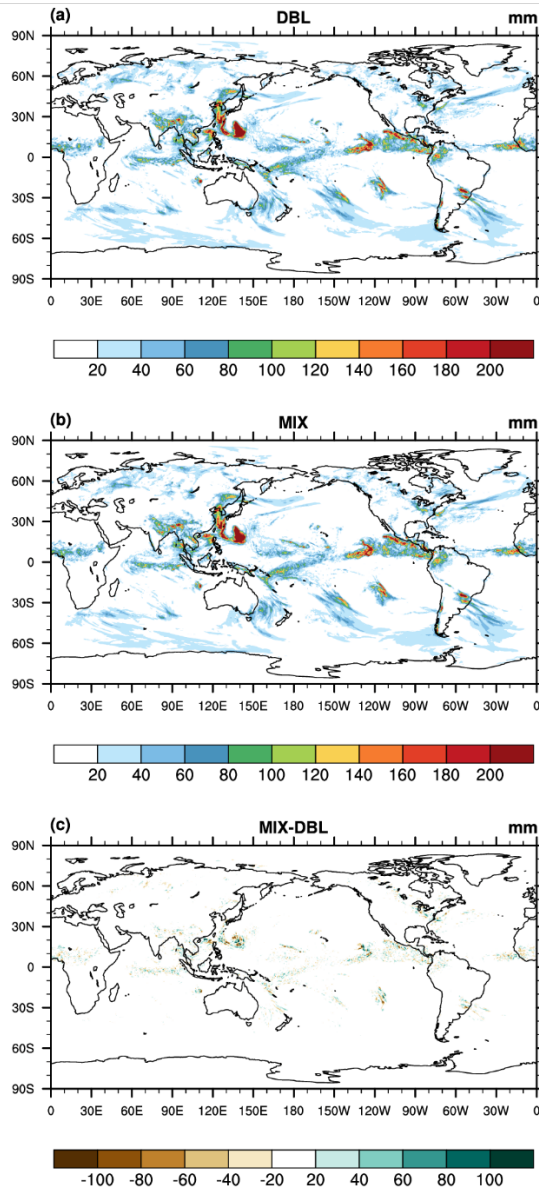
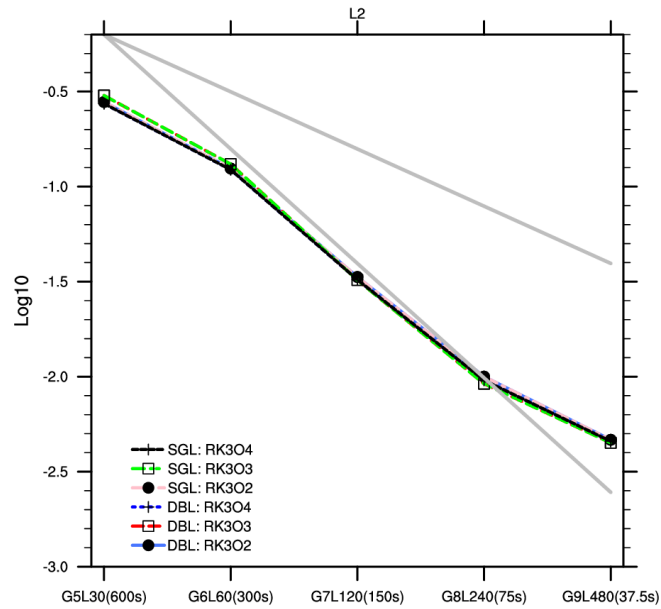



Figure 10: The 5-day (from 0000 UTC on 10 to 0000 UTC on 15 July 2015) accumulated precipitation (units: mm) from the (a) DBL simulation, (b) MIX simulation and (c) the difference between MIX and DBL simulations.

650 **Figure 11: The L_2 norm error and convergence rate of 3D passive transport test (Hadley-like meridional circulation) using single and double precision for three horizontal flux operators: RK3O2, RK3O3, and RK3O4. The upper and lower grey lines correspond to the slopes of first- and second-order convergence rates, respectively.**