Implementing and Evaluating Reduced-precision Calculations in EAMv1

Shixuan Zhang[1], Hui Wan[1], Phil Rasch[1], Balwinder Singh[1], Vince Larson[2], and Carol Woodward[3]

[1]Pacific Northwest National Laboratory, Richland, WA, USA
[2]University of Wisconsin—Milwaukee, Milwaukee, WI, USA
[3]Lawrence Livermore National Laboratory, Livermore, CA, USA

PNNL is operated by Battelle for the U.S. Department of Energy

Atmosphere Model, Whole Atmosphere, and Chemistry Climate Working Group Meeting, NCAR, Boulder, 9-11 March 2020

Background

- Reduced precision: benefits and challenges

Our method for implementing reduced-precision calculations and assessing their correctness

- The idea and its mathematical basis
- Validation of the method using E3SM
    - Idealized dynamical core tests
    - Simulations using mixed precision in the physics package

Conclusions and Future work

Motivation

- Current operation weather and climate models all use double-precision (64-bit) floating-point arithmetic
- Benefits of lower precision: fewer bits to calculate and communicate among parallel processes → lower computational cost *(e.g. Düben et al., 2014, 2015, 2017)*
- Is double-precision necessary for all variables in all calculations of an atmosphere model (i.e. the dynamical cores and parametrization)? *(Palmer et. al., 2014)*
- A series of recent studies have demonstrated that satisfactory simulation quality can be achieved with lower-precision arithmetic
    - Lorenz' 96 model *(e.g. Düben et. al., 2014)*
    - Dynamical cores of global atmospheric models *(e.g. Düben et. al., 2015)*
    - General circulation models *(e.g. Düben et. al., 2015; Jeffress et. al., 2017; Nakano et al., 2018)*

ECMWF’s Pioneering Work

- **Single-precision ensemble forecast** (*Váňa et al., 2017)*
  - Computational cost of a single realization (forecast) was reduced by ~40%
  - Investing the 40% on more members led to better ensemble forecasts
- **It was not trivial to make the single-precision configuration work**
  - Many pieces of the IFS code were not written with lower precision in mind
  - A small amount of calculations were found to need double precision
  - See Váňa et al. (2017, MWR) for code changes in IFS

List of code changes in IFS
(Váňa et al., 2017, MWR)

496 MONTHLY WEATHER REVIEW VOLUME 145

2. Code modifications allowing the use of single precision

The version of IFS used for the model simulations was based on model cycle CY41R2 (operational cycle at ECMWF since March 2016). The presented simulations are not coupled to an ocean model, though coupling to the operational wave model is included.

Technically the switch between single- and double-precision arithmetic is realized through the FORTRAN KIND parameter of all real variables in the code. This KIND parameter is defined in one module to modify the precision of all real variables. Additional code changes are required to allow both single- and double-precision execution with the same code. Those rather technical changes were as follows:

- Replace hard coded thresholds and security constants by parameters defined with intrinsic functions of precision [e.g., to replace code such as "x = 10.E+100" by an intrinsic FORTRAN function "x = huge(x)"].
- Extend code interfaces with I/O, MPI, and mathematical libraries (LAPACK/BLAS) to handle both double- and single-precision data.
- Ensure system functions (such as the system clock) are independent of the requested model precision.
- Make sure binary input files are read with the appropriate precision level.

To avoid run-time problems (i.e., floating-point exceptions or under/overflow), the model code had to be further analyzed to ensure all of the computations remain within the floating-point interval available to single precision (between 2^{-126} and 2^{127}). These changes were mostly seen as desirable in any case in terms of securing the general performance of the model, regardless of its actual precision. Typical changes in this respect were as follows:

- Increase security by avoiding divisions by floating-point numbers that may become zero.
- Rewrite a few formulas to make sure that all intermediate results stay within the single-precision range [e.g., rewrite A^{1}/B^{1} into $A(A/B)^{1}$].

Finally, to ensure that the single-precision model delivers meteorologically similar results to the double-precision reference, three parts in the model code had to be further modified.

- The cost of the radiation scheme is reduced by recomputing the fluxes on a longer time step than the model step. The contribution to a standard time step has to be upscaled from this coarser temporal sampling. This procedure was found responsible for introducing of an error for low solar zenith angles (below 86°) in the shortwave radiation code. The double-precision version of the code using a lower threshold to truncate numerical noise was in this case responsible for amplifying a higher amount of noise. This led to the difference of results from the two precision options. By rewriting the code to an equivalent but simpler formula, substituting the implicit and zero-division protected double division of the same variable by single multiplication (not requiring a protection), equal and improved performance for both precision levels was achieved.
- The setup for Legendre transformations, mainly the calculations of the roots and the associated polynomials,

FEBRUARY 2017 VÁŇA

is very sensitive to round-off errors due to their recurrence relations. Before the Swarztrauber algorithm (Swarztrauber and Spotz 2000) in the Legendre transforms was implemented in the IFS, this part of the model had to be computed with quadruple precision for resolutions higher than ~40km. When this code is run in single-precision arithmetic, the model results for higher resolution are substantially degraded. The solution adopted was to ensure that this code is always evaluated in double precision during the model setup (first time step). The Legendre transformation during the standard time step could then be evaluated in single precision.
- Use of single precision in the vertical finite element scheme (VFE) (Untch and Hortal 2004) was found responsible for amplifying model error through the pressure gradient term evaluation. This resulted in unacceptably poor mass conservation. The adopted solution, which almost entirely solved this issue, was to precompute the VFE integral operator with double-precision arithmetic. The resulting operator was truncated to a chosen precision.

The code changes outlined above have been regarded as beneficial with respect to code quality, an improved computational efficiency, and stability of the model. The exercise of exploring use of single precision exposed various problematic places in the code that could potentially also degrade the forecast when using double precision.

3. Results

For E3SM and similar (climate) models

- How to judge if a simulation using reduced precision gives correct results?
- When the results are incorrect, how to quickly identify problematic code pieces?

Assessing Solution Quality

For **weather models**, quantitative and objective prediction skill metrics are routinely used for model evaluation (i.e., their "score cards")

For **climate models**, a typical evaluation of model fidelity involves computing a large collection of statistical (climatological) features

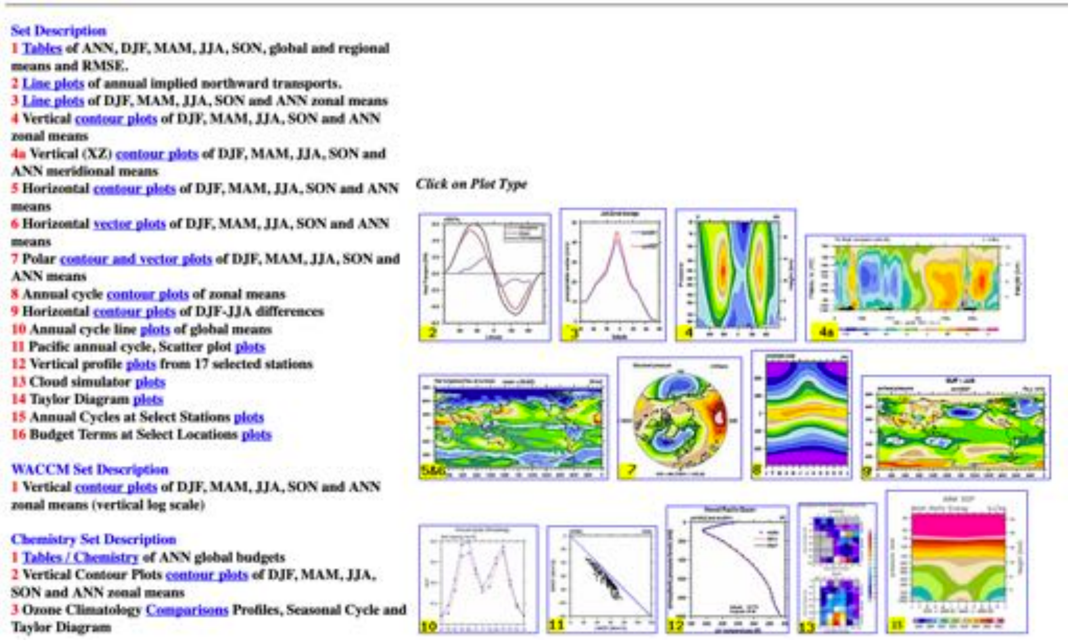
- Can be computationally expensive
- Lack of a concise overall fidelity metric

Data requirement of the AMWG diagnostics package:

- Minimum: 1 year
- To assess statistical significance: > 10 years

Package contains

- Order 100 to 1000 figures to review



For high-resolution models, it can be too expensive (or impractical) to conduct many 1-year simulations.

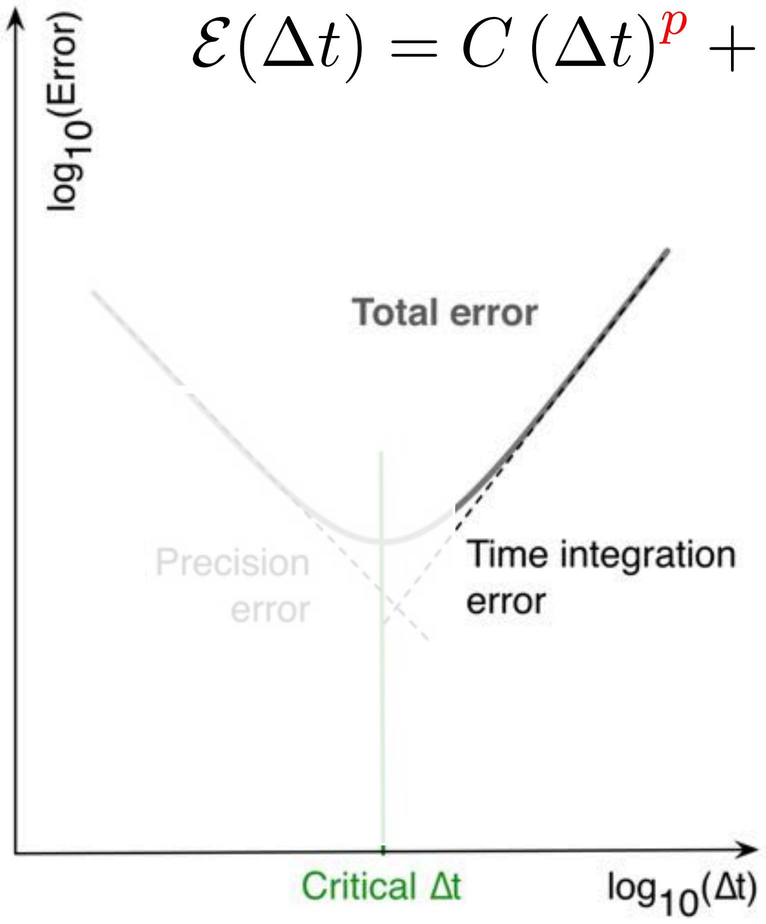
Our Proposed Approach to a Single-precision E3SM

- Incremental implementation
  - I.e., one parametrization at a time, or even one piece of a parameterization at a time
- Testing results using short simulations
- Assessing impact of reduced precision based on convergence behavior w.r.t. time step size

Precision Error v.s. Time-stepping Error

The Mathematical Basis

$$\mathcal{E}(\Delta t) = C\,(\Delta t)^{p} + \mathcal{R}$$



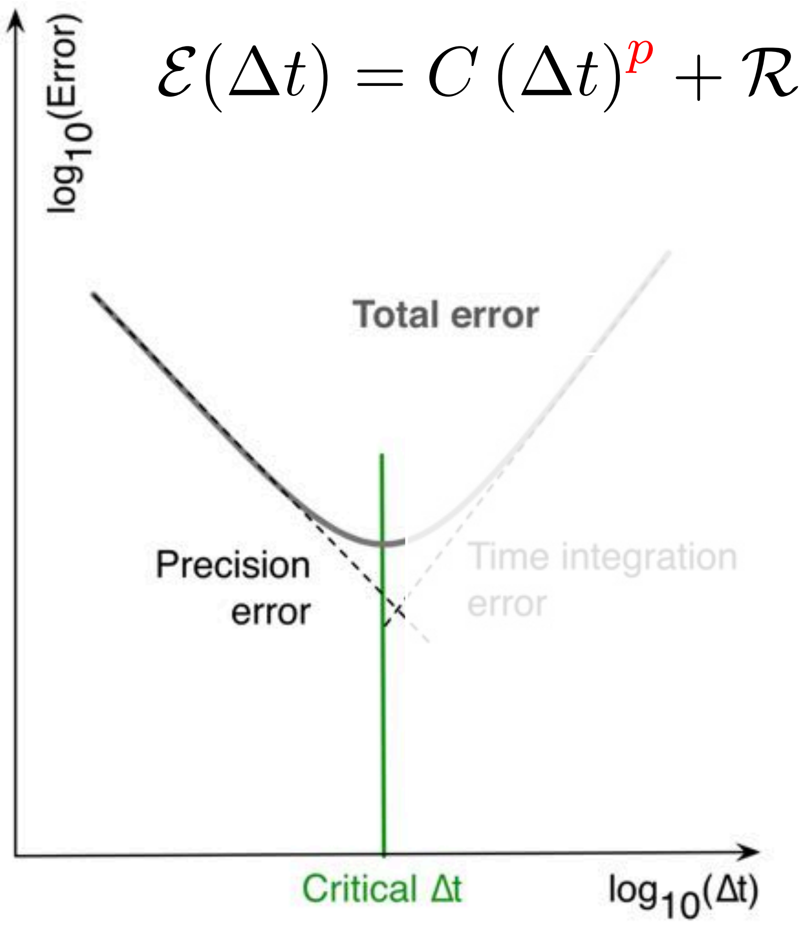
Zhang et. al. (2019, JAMES)
doi: 10.1029/2019MS001817

- For a numerically convergent discrete model, time-stepping error is expected to decrease when time steps are shortened

Precision Error v.s. Time-stepping Error

The Mathematical Basis

$$\mathcal{E}(\Delta t) = C\,(\Delta t)^{p} + \mathcal{R}$$



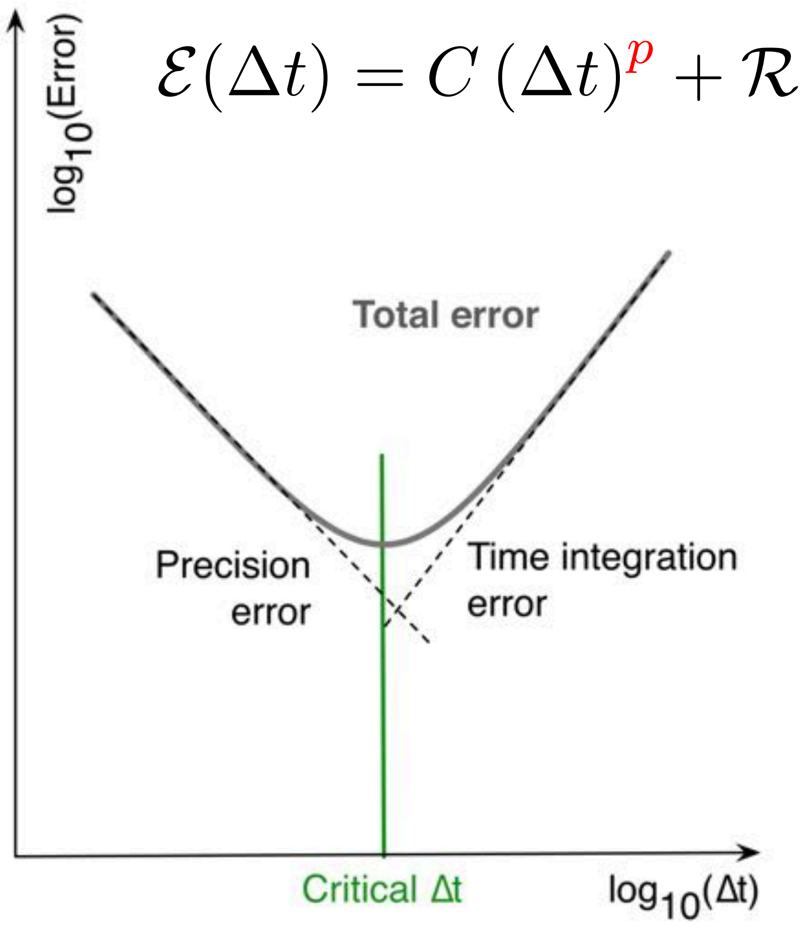
Zhang et. al. (2019, JAMES)
doi: 10.1029/2019MS001817

- When precision error dominates, smaller step sizes are expected to result in larger error because more steps are needed to finish a fixed-length simulation and hence allowing more floating-point operations to accumulate rounding error

Precision Error v.s. Time-stepping Error

The Mathematical Basis

$$\mathcal{E}(\Delta t) = C\,(\Delta t)^{p} + \mathcal{R}$$



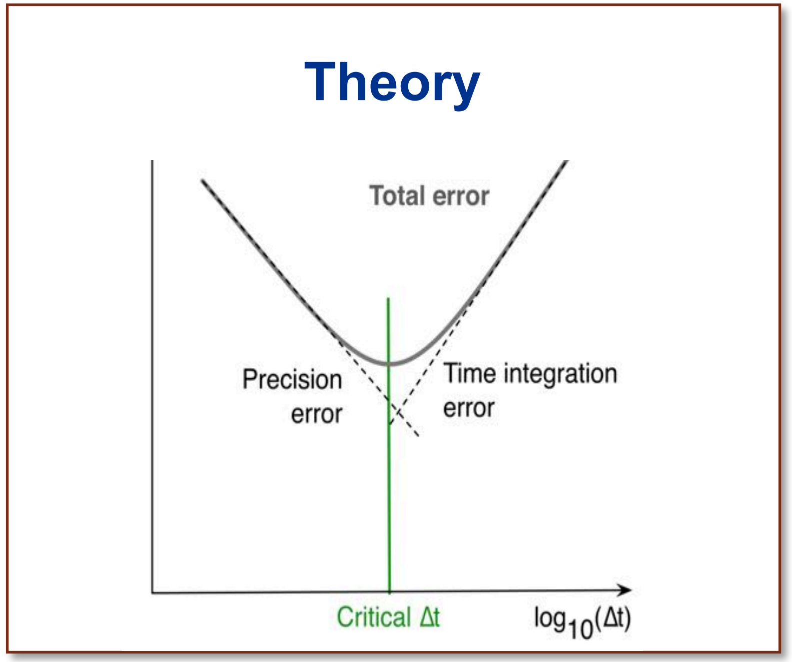
Zhang et. al. (2019, JAMES)
doi: 10.1029/2019MS001817

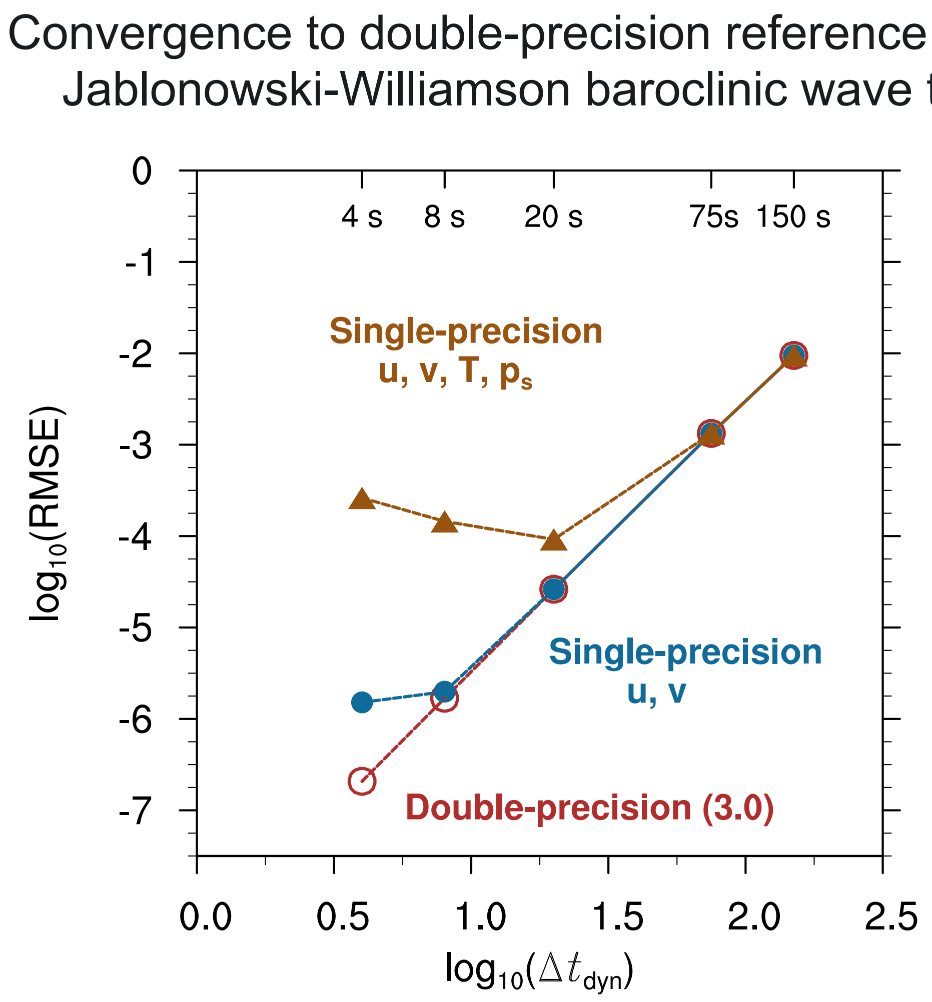
- The slope/shape of the convergence curve provides information about the relative magnitudes of precision error and time-stepping error

From Theory to Complex model

Idealized dynamical core test

Single precision emulated using code from Dawson and Düben (2017, GMD)





Zhang et. al. (2019, JAMES)
doi: 10.1029/2019MS001817

Simulation setup

- SE dynamical core only
- 3rd order configuration
- Jablonowski-Williamson test case
- 1-h global simulation
- Δt range: 150s down to 1s
- reference solution: double-precision with 2s

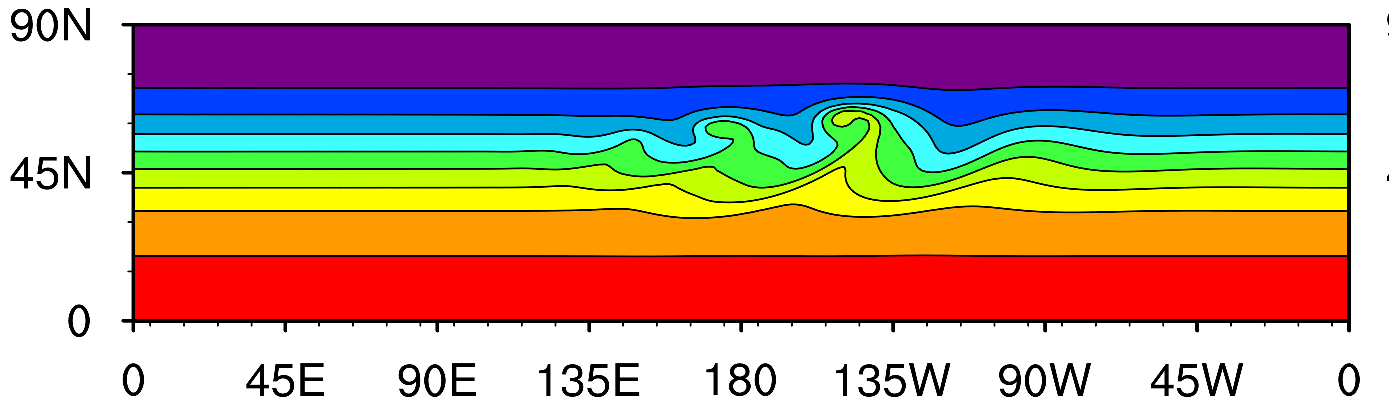
- SE dynamical core behaves as expected

Math vs Physics

- Essence of the method: Comparing precision error with time-stepping error to determine its significance ➔ math perspective
- Does the method tell us anything about the physics/fluid dynamics we care about? The answer is yes

Reduced Precision in SE dynamical core

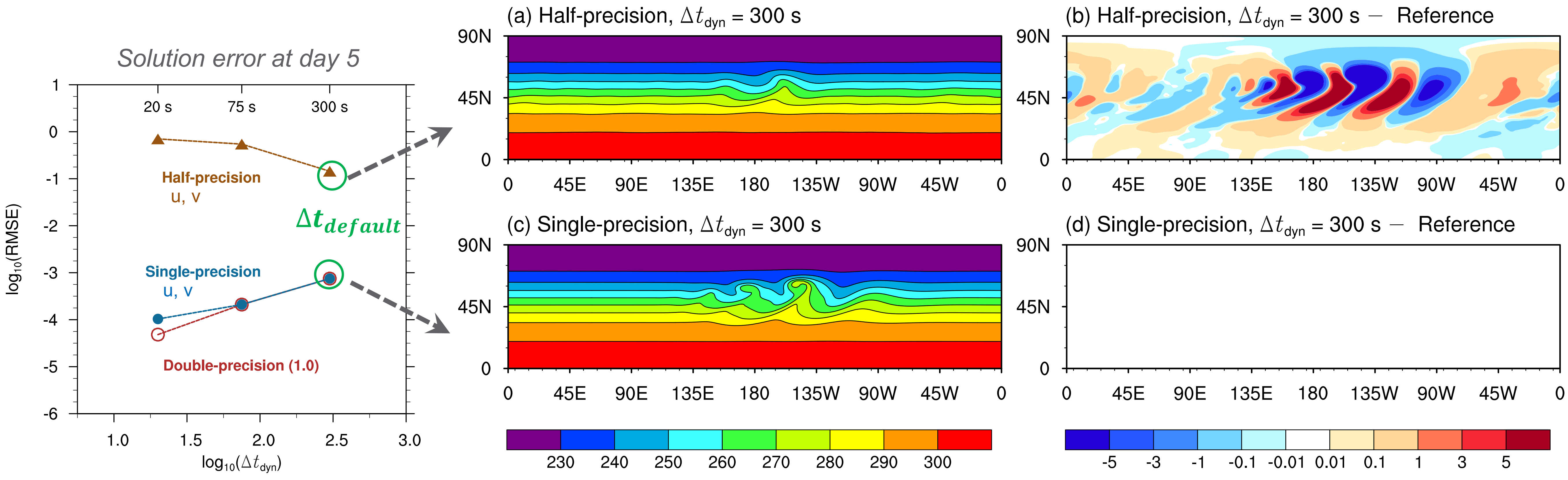
- Jablonowski-Williamson Dry Dynamical Core Test



Evolution of an idealized baroclinic wave in the northern hemisphere at day 9

- Two different levels of precision errors introduced to horizontal winds (u, v):
  - Single precision
  - Emulated half precision (Dawson and Düben, 2017, GMD)

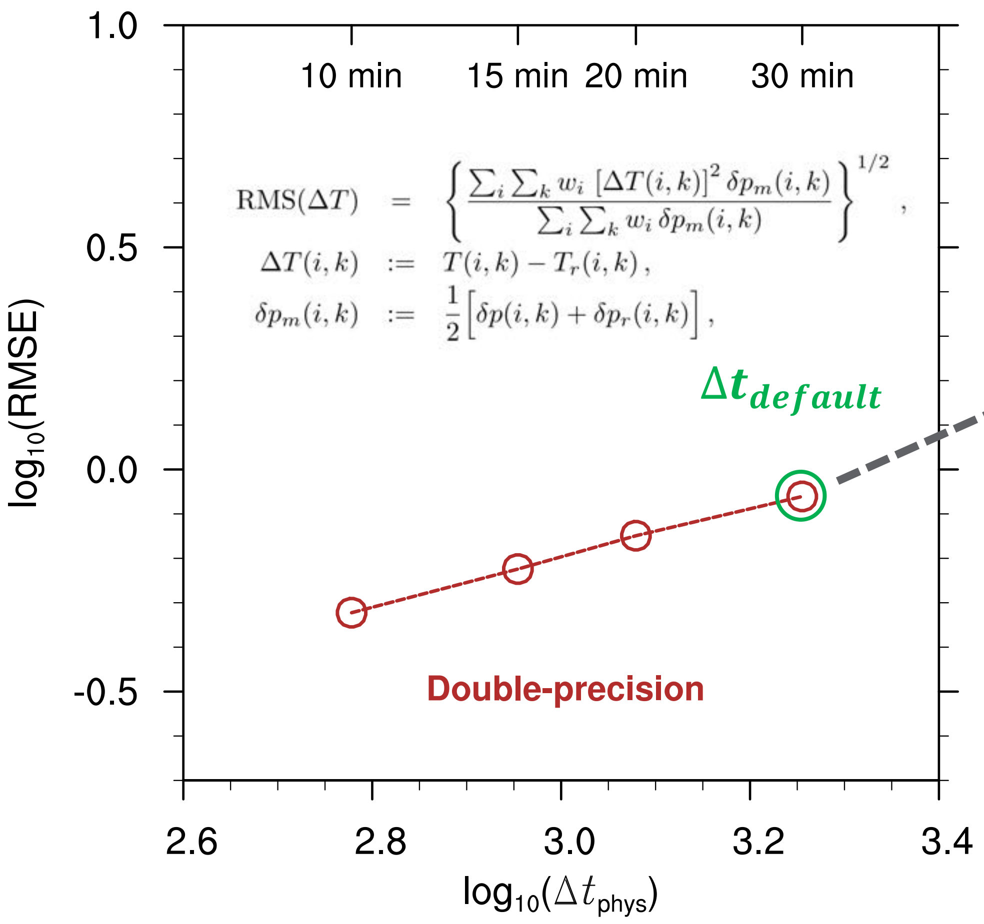
Solution Differences in Jablonowski-Williamson Test



The Jablonowski and Williamson (2006) baroclinic wave features at day 9 simulated by the dynamical core of EAMv1 using double, single or half precision (Zhang et al., 2019, doi: 10.1029/2019MS001817)

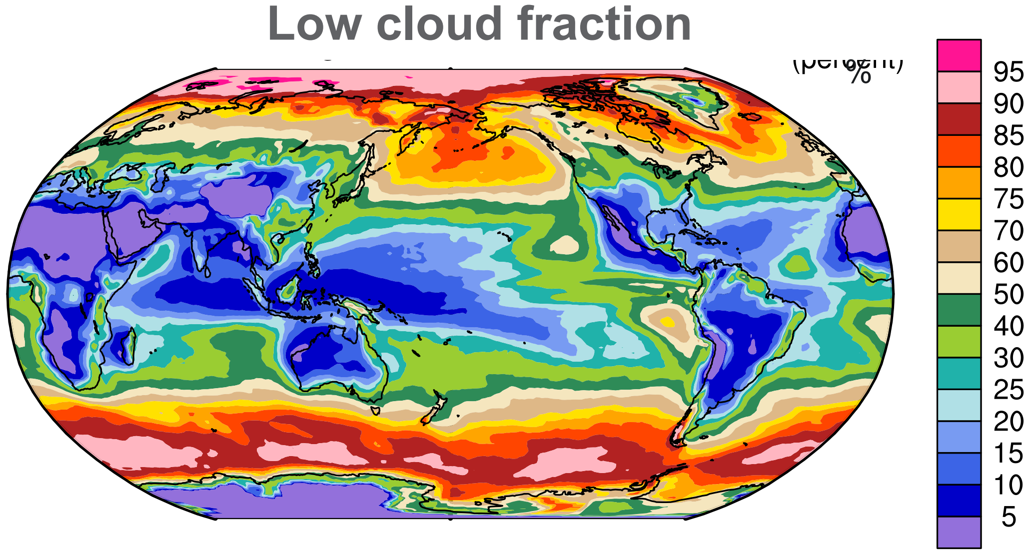
Simulations with Full Physics: Double Precision

5-day Convergence Test



Zhang et. al. (2019, doi: 10.1029/2019MS001817)

Evaluation using the traditional method: 10-year climate simulations

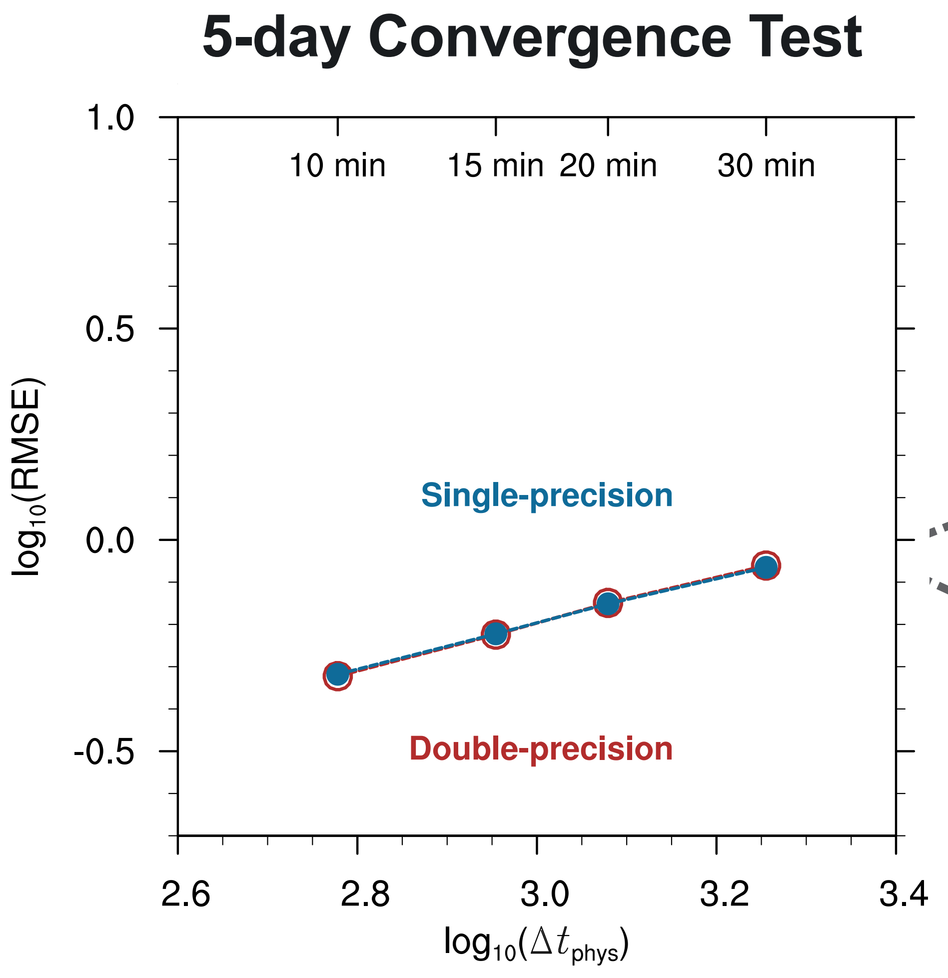


Climate simulation
- Default EAMv1
- FC5 compset
- 1-degree resolution

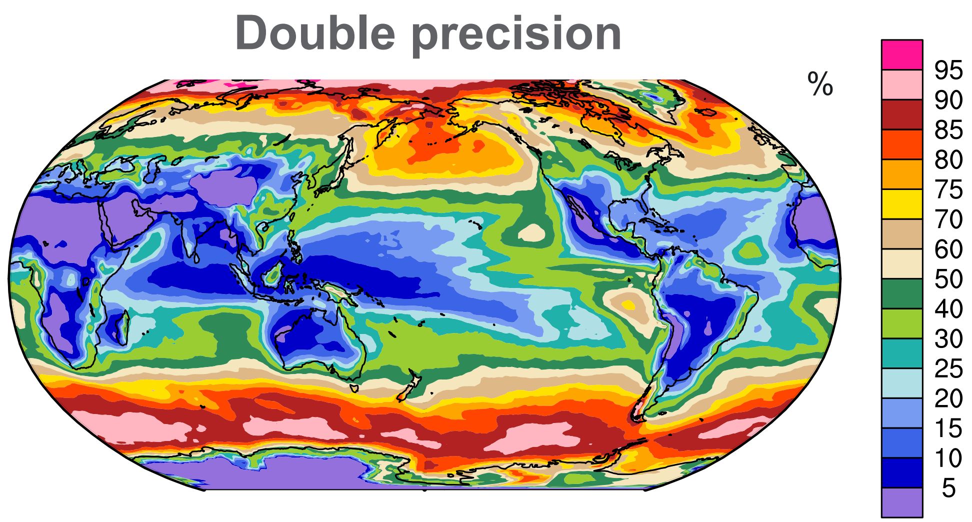
Convergence test
- 5-day global simulations
- Δt range: 30 min down to 5min
- reference solution: double-precision with 5min
- computing cost: 120 times cheaper than the traditional method with a 10-year simulation

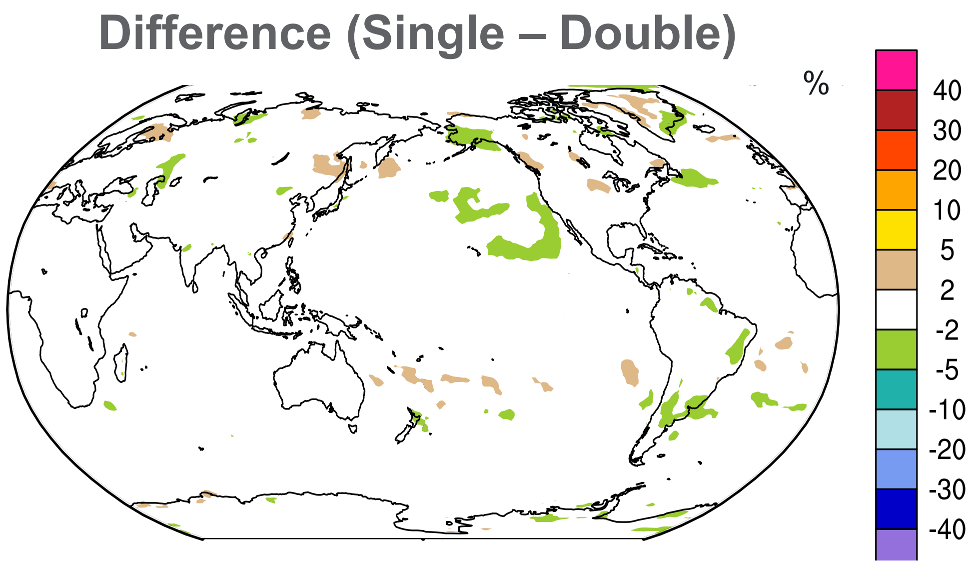
Mix-precision Configuration #1

- Single precision for CLUBB
- Double precision for rest of model



Evaluation using the traditional method: 10-year mean low cloud fraction

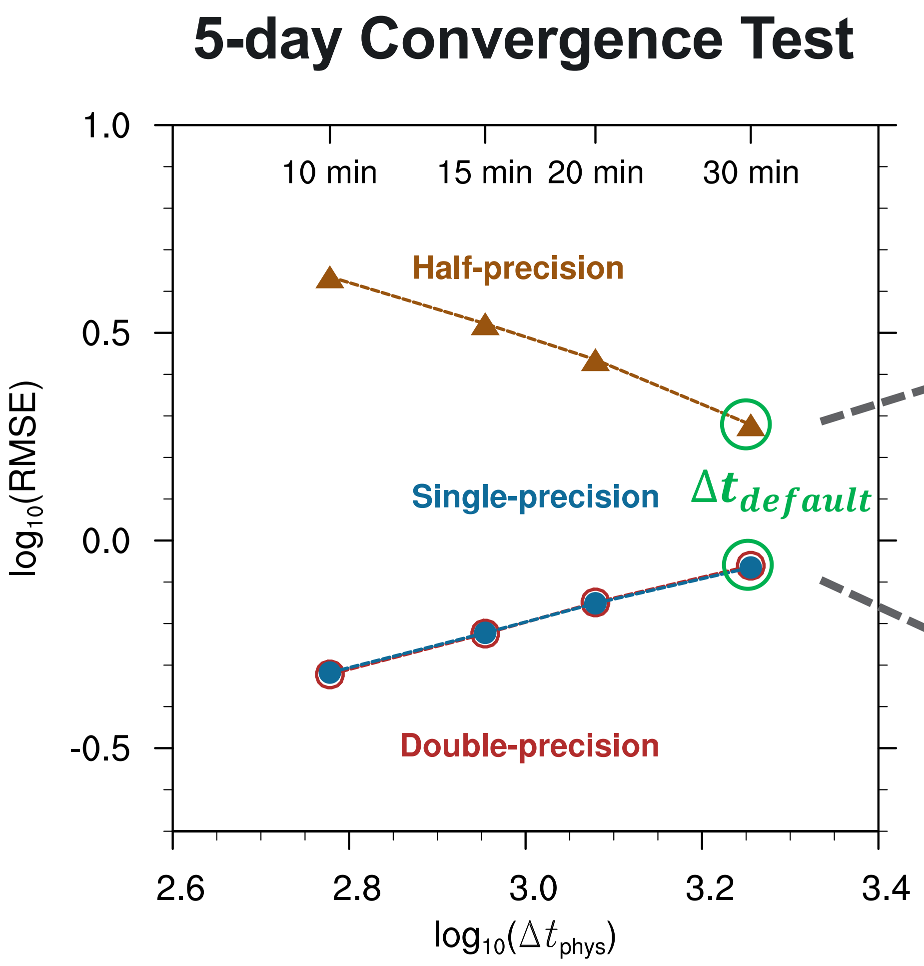




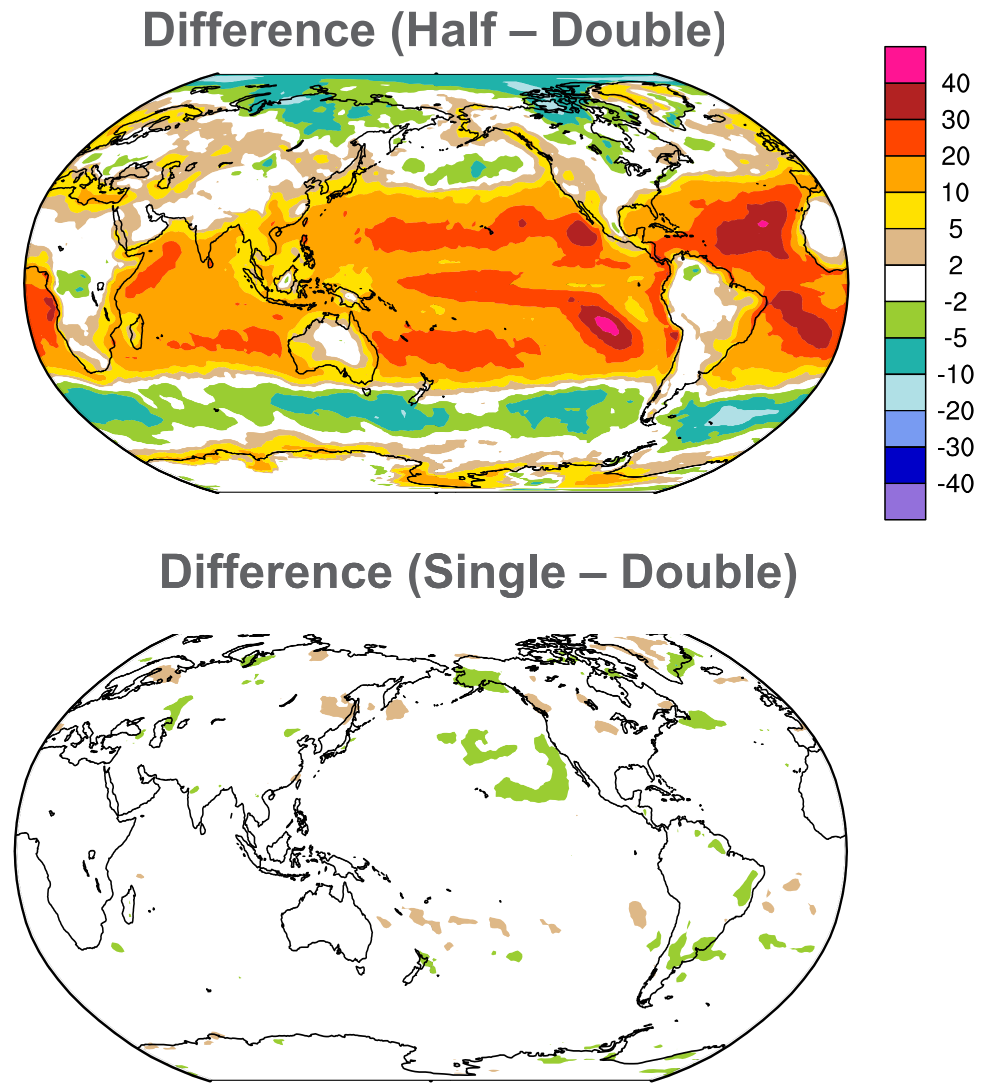
Zhang et. al. (2019, doi: 10.1029/2019MS001817)

Mix-precision Configuration #2

- Emulated **half** precision for CLUBB
- Double precision for rest of model

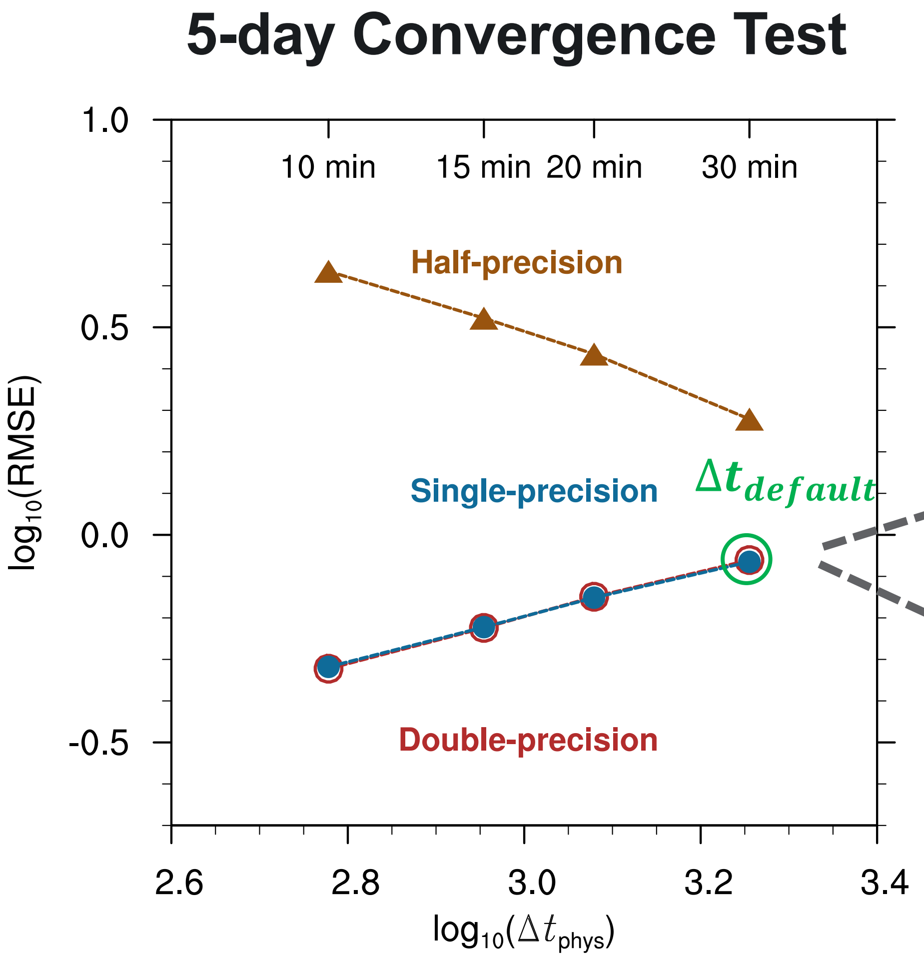


Evaluation using the traditional method:
10-year mean low cloud fraction

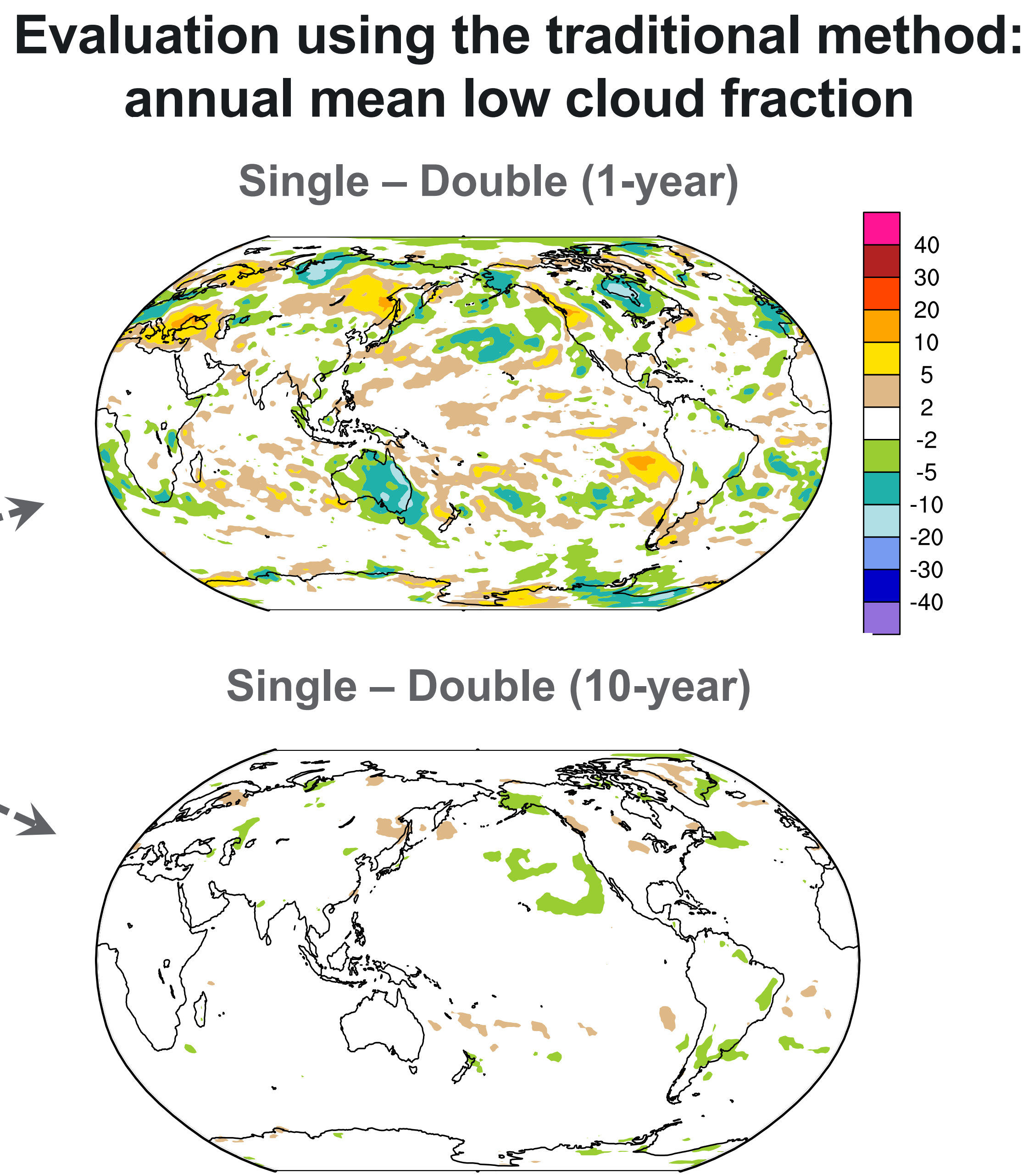


Zhang et. al. (2019, doi: 10.1029/2019MS001817)

Traditional Method vs. Convergence-based Method



Zhang et. al. (2019, doi: 10.1029/2019MS001817)



Summary

- We demonstrated a proof of concept that a single-precision E3SM is possible
- A simple, quantitative and objective method is developed to evaluate the correctness of model results
  - Short simulations successfully predict the impact of reduced precision on long-term climate features affected by fast physics
  - The method is computationally inexpensive, making it particularly useful for the development of higher-resolution models
- By incrementally converting more code pieces to single precision, we can eventually get a single-precision E3SM

Next Steps

- **EAGLES project**
  - Developing parameterizations of aerosols and aerosol-cloud interactions for the convection-permitting version of the E3SM atmosphere model
  - We will take this opportunity to implement a single-precision option for aerosol-related parameterizations

- **Further development of the test method**
  - Implement and evaluate the method in box-model and single-column simulations
  - Further consider some technical details, e.g.
    - How to choose simulation length for different physical processes?
    - Can we consolidate this precision error test with the solution reproducibility test of Wan et al. (2017, GMD)?

Backup slides

Beyond Single precision

- **Variable precision** model *(Palmer et al., 2014)*
  - E.g., processes of smaller scales are more uncertain, so there is no need for high precision
  - Smart idea
  - Challenging task (because a lot of knowledge is needed to optimize the precision configuration for each variable and calculation in the code)

- **Other ongoing efforts to improve computationally efficient of ESMs**
  - Reduce complexity in parameterizations
  - Use more efficient numerical algorithms (e.g., for tracer transport)
  - Software level: memory usage, parallelization; new programming model
    Hardware level: e.g., GPUs

- Variable precision arithmetic + inexact computing hardware+ other efforts: >> 2x speed up