Mixed-precision for Linear Solvers in Global Geophysical Flows

Ackmann Jan¹, Dueben Peter Dominik², Palmer Tim N¹, and Smolarkiewicz Piotr K.³

¹Oxford University
²ECMWF
³National Center for Atmospheric Research

November 16, 2022

Abstract

Semi-implicit time-stepping schemes for atmosphere and ocean models require elliptic solvers that work efficiently on modern supercomputers. This paper reports our study of the potential computational savings when using mixed precision arithmetic in the elliptic solvers. Precision levels as low as half (16 bits) are used and a detailed evaluation of the impact of reduced precision on the solver convergence and the solution quality is performed.

This study is conducted in the context of a novel semi-implicit shallow-water model on the sphere, purposely designed to mimic numerical intricacies of modern all-scale weather and climate (W&C) models. The governing algorithm of the shallow-water model is based on the non-oscillatory MPDATA methods for geophysical flows, whereas the resulting elliptic problem employs a strongly preconditioned non-symmetric Krylov-subspace solver GCR, proven in advanced atmospheric applications. The classical longitude/latitude grid is deliberately chosen to retain the stiffness of global W&C models. The analysis of the precision reduction is done on a software level, using an emulator, whereas the performance is measured on actual reduced precision hardware. The reduced-precision experiments are conducted for established dynamical-core test-cases, like the Rossby-Haurwitz wavenumber 4 and a zonal orographic flow.

The study shows that selected key components of the elliptic solver, most prominently the preconditioning and the application of the linear operator, can be performed at the level of half precision. For these components, the use of half precision is found to yield a speed-up of a factor 4 compared to double precision for a wide range of problem sizes.
Mixed-precision for Linear Solvers in Global Geophysical Flows

Jan Ackmann¹, Peter D. Dueben², Tim N. Palmer¹, Piotr K. Smolarkiewicz³

¹University of Oxford, Oxford, OX1 3PU, UK
²European Centre For Medium Range Weather Forecasts, Reading, RG2 9AX, UK
³National Center for Atmospheric Research, Boulder, CO 80026, USA

Key Points:

• A detailed study of mixed-precision for linear solvers for weather and climate models is performed, including half precision with 16 bits
• Precision can be reduced in important parts of the solver but a naive approach to reduce precision everywhere does not work
• Compute time can be reduced by a factor four compared to double precision for parts that work in half precision

Corresponding author: Peter D. Dueben, peter.dueben@ecmwf.int
Abstract

Semi-implicit time-stepping schemes for atmosphere and ocean models require elliptic solvers that work efficiently on modern supercomputers. This paper reports our study of the potential computational savings when using mixed precision arithmetic in the elliptic solvers. Precision levels as low as half (16 bits) are used and a detailed evaluation of the impact of reduced precision on the solver convergence and the solution quality is performed.

This study is conducted in the context of a novel semi-implicit shallow-water model on the sphere, purposely designed to mimic numerical intricacies of modern all-scale weather and climate (W&C) models. The governing algorithm of the shallow-water model is based on the non-oscillatory MPDATA methods for geophysical flows, whereas the resulting elliptic problem employs a strongly preconditioned non-symmetric Krylov-subspace solver GCR, proven in advanced atmospheric applications. The classical longitude/latitude grid is deliberately chosen to retain the stiffness of global W&C models. The analysis of the precision reduction is done on a software level, using an emulator, whereas the performance is measured on actual reduced precision hardware. The reduced-precision experiments are conducted for established dynamical-core test-cases, like the Rossby-Haurwitz wavenumber 4 and a zonal orographic flow.

The study shows that selected key components of the elliptic solver, most prominently the preconditioning and the application of the linear operator, can be performed at the level of half precision. For these components, the use of half precision is found to yield a speed-up of a factor 4 compared to double precision for a wide range of problem sizes.

Plain Language Summary

Numerical models of the Earth system are very important to predict weather and climate. The models are computationally expensive and run on supercomputers. If the models are enabled to run faster, model simulations can be enhanced – for example via a higher resolution – and the accuracy of predictions can be improved.

One way to enable the models to run faster is via the use of very low numerical precision. For example via the use of so-called half precision which is representing all variables of the simulation with only 16 bits, instead of the default value of 64 bits, reducing computational time by up to a factor of four. In this paper, we investigate whether numerical precision can be reduced down to half precision for parts of the linear solvers which are used in some weather and climate models. Linear solvers help to progress the model state in time and are responsible for a large fraction of the overall computational cost of model simulations.

1 Introduction

In the simulation of complex multi-scale flows arising in weather and climate (W&C), one of the biggest challenges is to satisfy strict service requirements in terms of time-to-solution and to satisfy budgetary constraints in terms of energy-to-solution, without compromising the accuracy and stability of the application (A. Müller et al., 2019). One way to tackle this challenge is to use reduced-precision arithmetic in W&C models. The interest with reduced precision or the combination of different precision levels in "mixed precision" approaches has already some history in scientific computation (Göddeke et al., 2007; Baboulin et al., 2009; Furnichi et al., 2011; Chen et al., 2012). However, it is relatively new in W&C modelling. This effort started with (Palmer, 2012) and (Düben et al., 2014) where reduced precision was motivated by the large degree of uncertainty that is present in W&C models due to their nonlinear dynamics and high level of com-
plexity which makes it difficult to justify high numerical precision. The study of mixed
precision in numerical modelling is timely, since the recent boom of machine learning meth-
ods is pushing developments of supercomputing hardware towards very efficient dense
linear algebra performed at low precision. This hardware is specialised for the use in deep
learning where half precision arithmetic—or even less—is often sufficient; see for exam-
ples the Tensor Processing Unit (TPU) by Google.

A large part of W&C models is used to represent subgrid scale processes—e.g., tur-
bulence and cloud physics—that carry a large degree of inherent uncertainty (Saffin et
al., 2020). In principle one can thus argue that representation of such processes is the
prime candidate for a precision reduction. The other dominant part of W&C models is
the dynamical core (a.k.a dycore) which provides discrete solutions to the 
Navier-Stokes
equations. Although there is little uncertainty about the equations per se, their numer-
ical solution procedures introduce model error that add uncertainty to the dynamics. Early
work on thorough analysis to identify the minimal level of numerical precision that can
be used in different parts of dycores has focused on spectral model formulations (Düben
& Palmer, 2014; Chantry et al., 2019). Similar work on grid-point model formulations
has only just begun with the use of single precision arithmetic (Nakano et al., 2018; Tintó Prims
et al., 2019; Maynard & Walters, 2019) with 32 bits per variable.

In efficient W&C dycores the semi-implicit (SI) timestepping is typically used for
the atmospheric (Mengaldo et al., 2019) as well as the oceanic (Adcroft et al., 2004; Korn,
2017; Wang et al., 2014) component. Its key virtue is extended computational stability
with respect to acoustic, buoyant and rotational modes of motion allowing for the use
of relatively long time steps. On the other hand, the approach results in an intricate lin-
ear problem and complicated elliptic boundary value problem for pressure (viz. Schur
complement). In particular, the SI approach is at the heart of the newly-developed Finite-
Volume Module (FVM) of the Integrated Forecasting System (IFS) at the European Cen-
tre for Medium-Range Weather Forecasts (ECMWF) (Smolarkiewicz et al., 2016, 2017,
2019; Kühnlein et al., 2019). The FVM provides an alternative dynamical core to the
spectral transform based IFS. The two essential ingredients of the FVM timestepping
is the Multidimensional Positive Definite Advection Transport Algorithm (MPDATA)
approach (Smolarkiewicz & Szmelter, 2005; Szmelter & Smolarkiewicz, 2010; Kühnlein
& Smolarkiewicz, 2017) and the bespoke preconditioned nonsymmetric Krylov-subspace
elliptic solver (Kühnlein et al., 2019) built on the Generalized Conjugated-Residual (GCR)
approach of (Eisenstat et al., 1983). While MPDATA controls the explicit part of the
timestepping, GCR inverts the Schur complement of the linear problem.

Undeniably, elliptic solvers are computationally demanding and substantial devel-
opment has been invested into making them as efficient as possible for W&C models on
modern supercomputers (E. H. Müller & Scheichl, 2014; Yang et al., 2016; Kühnlein et
al., 2019). To efficiently solve an elliptic problem posed in a thin spherical shell (such
as the global atmosphere) ultimately requires matrix inversion—if not in the main solver
than at least in its preconditioner. While there is the common opinion that high preci-
sion is required for this purpose, there are theoretical studies (Göddeke et al., 2007; Baboulin
et al., 2009; Carson & Higham, 2018; Haidar et al., 2018; Anzt et al., 2019) suggesting
that mixed-precision approaches could be exploited, and there are already some appli-
cations of mixed-precision elliptic solvers in CFD (Idomura et al., 2018; Amritkar & Tafti,
2016; Furuichi et al., 2011) and also in W&C (Maynard & Walters, 2019). Although these
approaches may differ in choice of algorithms and applications, a common theme emerges
that the preconditioning step may be a good choice for reducing precision, where some
work goes as low as half precision arithmetic representing each real number with only
16 bits (as a floating point number) and decimal precision reduced to three digits.

The aim of this paper is to establish the limits of precision for the specific combi-
nation of characteristics that make SI W&C models unique, while its objective is to per-
form an in-depth analysis of the behavior of mixed-precision elliptic solvers in such an
environment. The elliptic problems encountered in SI W&C models are non-symmetric, and are typically solved using iterative solvers due to the enormous problem size. We solve a fluid dynamics problem on the sphere, which comes with grid irregularities, giving rise to a specific structure of the involved linear operator. In effect, a standard classification of the problem based solely on the operator condition number does not give full justice to the complexity of the problem; cf. (Smolarkiewicz et al., 2019; Kühnlein et al., 2019). Moreover, the elliptic problem is part of an entire SI timestepping scheme and as such is in a feedback loop with the non-linear dynamics of the model. Consequently, we solve much more at each time step than a sole elliptic problem. Different solver solutions, even obtained to the same accuracy in terms of the residual errors, can lead to vastly different behavior and model error growth throughout a simulation. For these reasons, our mixed-precision elliptic solver needs to be tailored to the specific application at hand and thoroughly tested in this specific W&C environment.

Herein, we study the impact of mixed-precision in the elliptic solver for W&C models by investigating a representative SI shallow-water model on the sphere and using the MPDATA approach. With these choices, we stay conceptionally close to the ideas employed in the IFS-FVM. The latter is formulated in the latitude-longitude (lat-lon) coordinate framework, but circumvents polar grid singularity via a quasi-uniform unstructured mesh discretisation, whereby the essential problem stiffness comes from the thin vertical dimension. The shallow-water model of this paper uses a regular lat-lon grid subject to polar singularity (Prusa, 2018), but retains the aspect of the problem stiffness which is coming from the longitudinal dimension. Given the stiffness and grid singularity that are already prone to creating model errors in double precision arithmetic (64 bits per variable), we carefully study what happens to solutions in polar regions under reduced precision. We test our mixed-precision solver on test-cases from the well-known test-suite for shallow-water dycores (Williamson et al., 1992). The first test-case is the standard problem of the Rossby-Haurwitz Wave with wave number 4. The second test-case paraphrases the standard zonal flow over an isolated mountain by replacing the idealised smooth hill with the natural orography of the Earth.

In this paper, whenever we refer to mixed-precision, we mean the use of different precision levels within the same simulation. This can either refer to the use of single and double precision as available on modern computers, or the use of half, single and double precision. While half precision is available on machine learning accelerators on modern supercomputing hardware, it is currently not yet available for use in most high-level computer languages. An important milestone in this regard is that Fugaku, the current number one of the TOP500 list of supercomputers, allows for the use of half precision arithmetic from Fortran. For this paper, half precision will be emulated in software as well as used on actual hardware, here on the Fujitsu A64FX processor available on the supercomputer Isambard 2. The reduced-precision emulator is used because it enables to easily explore different precision levels, while the ensuing performance measurements are performed on Isambard 2.

The paper is structured as follows. In Section 2, we describe the continuous as well as discretised model and introduce our elliptic solver and the preconditioner. Section 3 presents the test-cases, the model setup and a description of the reduced-precision experiment suite. In section 4 we show and discuss reduced-precision results. Section 5 concludes the paper.

2 Model Description

2.1 The Governing Equations

With our choice of model, we aim to stay conceptionally as close as possible to the approach used by other MPDATA based fluid models (Prusa et al., 2008; Szmelter & Smo-
larkiewicz, 2010; Kurowski et al., 2016) and consequently the newly developed IFS-FVM atmospheric model (Smolarkiewicz et al., 2016, 2019; Kühllein et al., 2019). Thus, we chose to use the shallow-water equations (Smolarkiewicz & Margolin, 1998; Szmelter & Smolarkiewicz, 2010):

\[
\frac{\partial G\Phi}{\partial t} + \nabla \cdot (v\Phi) = 0, \quad (1)
\]

\[
\frac{\partial GQ_x}{\partial t} + \nabla \cdot (vQ_x) = GR_x, \quad (2)
\]

\[
\frac{\partial GQ_y}{\partial t} + \nabla \cdot (vQ_y) = GR_y, \quad (3)
\]

where \( \Phi \) is the fluid thickness, \( \nabla = (\partial/\partial x, \partial/\partial y) \), \( Q_x = \Phi \dot{x}h_x \) and \( Q_y = \Phi \dot{y}h_y \) denote the momenta in \( x \) (longitudinal) and \( y \) (latitudinal) directions, while \( G = h_xh_y \) is the Jacobian of the geospherical framework with \( h_x \), \( h_y \) being the metric coefficients of the general orthogonal coordinates. Here, \( h_x = a \cos(\phi) \), \( h_y = a \) for a lat-lon grid with Earth’s radius \( a \), so longitude \( \lambda = x \) and latitude \( \phi = y \). The advective velocity \( v \) relates to the momentum \( Q \) via

\[
v = G \left( \dot{x}, \dot{y} \right) = (Q_xh_y, Q_yh_x)/\Phi; \quad (4)
\]

cf. (Smolarkiewicz & Margolin, 1998; Szmelter & Smolarkiewicz, 2010) for discussions. The right-hand sides of the equations (2) and (3) symbolise the forcing terms \( R_x \) and \( R_y \) for the momenta, such that

\[
R_x = -\frac{g}{h_x} \Phi \frac{\partial (\Phi + H_0)}{\partial x} + fQ_y + \frac{1}{G\Phi} \left( Q_y \frac{\partial h_y}{\partial x} - Q_x \frac{\partial h_x}{\partial y} \right) Q_y - \alpha(Q_x - Q_y), \quad (5)
\]

\[
R_y = -\frac{g}{h_y} \Phi \frac{\partial (\Phi + H_0)}{\partial y} - fQ_x - \frac{1}{G\Phi} \left( Q_y \frac{\partial h_y}{\partial x} - Q_x \frac{\partial h_x}{\partial y} \right) Q_x - \alpha(Q_y - Q_x), \quad (6)
\]

with specific terms, from left to right, representing the pressure gradient, the Coriolis force, the metric forces (viz. Christoffel symbols of the geospherical framework) and the relaxation forcings attenuating the solution to some hypothetical “true” state \( \bar{Q} \) with the inverse time scale \( \alpha \). \( H_0 \) is the topography, \( g \) the gravitational acceleration and \( f \) the Coriolis parameter.

### 2.2 The Discretised Model

The below-summarised procedure of formulating an implicit linear problem for dependent model variables, together with its associated Schur complement (i.e. elliptic Helmholtz equation) is a special case of a substantially more intricate procedure documented recently for all-scale atmospheric Euler equations in (Smolarkiewicz et al., 2019), where the interested reader is referred for further details.

The governing Equations (1)-(3) are discretised in a semi-implicit fashion, using a collocated lat-lon grid with its well-known pole problem (Prusa, 2018). In a nutshell, momentum equations are integrated with the trapezoidal rule

\[
Q_{i}^{n+1} = M_1 \left( Q^n + 0.5\Delta t(R^n, v^{n+0.5}, G) + 0.5\Delta t\hat{R}_{i}^{n+1} \right) \equiv \hat{Q}_i + 0.5\Delta t\hat{R}_{i}^{n+1}; \quad (7)
\]

where \( n \) and \( i \) index discrete points (\( t^n, x_i \)) separated with uniform intervals \( \Delta t \) and \( (\Delta \lambda, \Delta \phi) \), respectively, in time and the two spatial directions. The first term on the rhs of (7) denotes a second-order MPDATA operator—with \( v^{n+0.5} \) symbolising an \( O(\Delta t^2) \) explicit predictor of advective velocity at the intermediate \( t^n + 0.5\Delta t \) time level; cf. §3.4 and §4.1 in (Smolarkiewicz & Margolin, 1998)—and it forms the explicit part of the solution. \( R^n \) is readily calculated as it is a function of known variables. The second term however is unknown, as \( R^{n+1} \) depends (nonlinearly) both on \( \Phi^{n+1} \) and \( Q^{n+1} \). There are sev-
eral alternative approaches to assure an easily invertable linear problem, involving outer
iterations or explicit predictors, or both (Smolarkiewicz et al., 2016, 2017, 2019; Kühnlein
et al., 2019). As our goal is to assess the role of mixed precision on the elliptic solver per-
formance, we here select an ad hoc approach that complicates (rather than simplifies)
the target elliptic solver.

Specifically, the pressure gradient term is approximated as

\[ \nabla H = 0.5[\nabla H(\Phi^* + H_0) + \nabla H(\Phi^{n+1} + H_0)] , \]

where \( \Phi^* \) denotes the explicit second-order-accurate predictor of \( \Phi^{n+1} \), generated by in-
tegrating the mass-continuity equation (1) with MPDATA. The Coriolis and relaxation
terms are implicit at \( t^{n+1} \), whereas the nonlinear metric forces—generally acting on a
much longer time scale than the scale associated with the propagation of external grav-
ity waves—are predicted explicitly at \( n + 1 \) by linear extrapolation from \( n \) and \( n - 1 \)
time levels.\(^1\) Thanks to the collocated grid, the resulting linear problem is inverted analytically, forming the closed-form expressions

\[ Q_i^{n+1} = \tilde{Q}_i + A_i \nabla \Phi_i^{n+1} + B_i \Phi_i^{n+1} , \]

where \( \tilde{Q} \), \( A \) and \( B \) denote, respectively, the modified explicit part of the problem, and
the \( 2 \times 2 \) matrix and vector of known coefficient fields. Implementing the expressions
(9) in the trapezoidal integral of (1), leads to the elliptic Helmholtz problem for \( \Phi^{n+1} \),
which can be symbolically written as

\[ \forall (n+1, i) \quad L_i(\Phi^{n+1}) - R_i^{n+1} = 0 ; \]

hereinafter, spatial and temporal grid indices are dropped as there is no ambiguity. The
linear operator \( L \) is negative definite but not self-adjoint. It takes the form of a gener-
ialized Laplacian

\[ L(\Phi) := \sum_{I=1}^{M} \frac{\partial}{\partial x^I} \left( \sum_{J=1}^{M} A^{IJ} \frac{\partial \Phi}{\partial x^J} + B^I \Phi \right) - C \Phi , \]

where the coefficient fields are straightforward modifications of those appearing in (9),
and \( M = 2 \). The coefficients for this shallow-water model are given in appendix Ap-
pendix A. The solution of (10) provides the updated \( \Phi \)—upon which \( Q^{n+1} \) is recovered
from (9) and then \( v^{n+1} \) from (4)—thus allowing to complete (7) for calculations with
large \( \Delta t \), limited only by the advective CFL condition.

2.3 The Elliptic Solver

The elliptic problem (10) is solved using the preconditioned generalized conjugate
residual (GCR) approach (Smolarkiewicz & Margolin, 2000; Eisenstat et al., 1983; Gillard
& Benacchio, 2022). GCR is a Krylov sub-space method that assures monotone conve-
gence of the solver iterations for a non-self-adjoint operator \( L \) (i.e., "non-symmetric" in
matrix representation) by minimising \( \langle rr \rangle \) (where \( \langle .. \rangle \) marks the domain integral), which
is equivalent to the Euclidean norm \( L_2(r) \) squared, of the residual error

\[ r_\nu = L(\Phi_\nu) - R , \]

here \( \nu \) numbers the iterations. In essence, GCR cleverly re-optimises at each iteration
all coefficients in the series of the current and past successive iterations up to some ar-
bitrary specified number, say \( k-1 \); cf. (Smolarkiewicz & Szmelter, 2011) for a discus-

\(^1\) The resulting integral of (1)-(3) is by construction second-order-accurate in time and space; cf. (Kühnlein et al., 2019) for a pertinent discussion.
sion. When after some number of iterations $N$, $L_2(r)$ reaches a desired small error tolerance, the solution of this iterative process $\Phi_N$ is set to be $\Phi^{n+1} = \Phi_N$. Figure 1 describes our implementation of the preconditioned GCR algorithm. Given the experience with the EULAG model (Prusa et al., 2008), GCR(3) is used.

Figure 1. From left to right: a flow chart of our preconditioned GCR(k) implementation; normalized runtimes of the four distinct algorithmic steps for GCR(3) measured in double precision (These performance measurements have been performend on a single core (CPU: Intel i5, Fortran compiler: GCC 9.3.0.) and they agree with the ratio of the number of occurring floating point operations); and the choice of base precision for each algorithmic step (The choices are described in more detail in §4.3).

Following the step 2 of the algorithm—the update of the depth $\Phi$ and the minimised residual error—there is an exit condition. While a variety of exit conditions can be considered (Smolarkiewicz et al., 1997), we found that demanding the uniform reduction of the initial residual error,

$$
\| r_{\nu+1} \|_2 \leq \epsilon \| r_0 \|_2 ,
$$

assures the stability of the semi-implicit model integrator while maintaining the solution quality throughout a range of spatial resolutions and physical scenarios. In practice, the choice of $\epsilon$ is verified experimentally. Based on our overall experience with W&C models, in this paper $\epsilon = 10^{-5}$ is assumed for all experiments. We also enforce at least one full GCR(3) cycle regardless of (13). Further discussion and justification for this choice of $\epsilon$ is found in §4.1.
The preconditioner $P$ in step 3 approximates $L$ indirectly, by estimating the solution error $q = P^{-1}(r) \approx e = L^{-1}(r)$ by means of a stationary iteration—indexed with $\mu$, such that $q_{\mu=0} = 0$—best characterised as a semi-implicit Richardson scheme (Smolarkiewicz & Margolin, 2000). Specifically, the operator $P$ is split in two parts. The first part combines the second-order zonal derivative term and the Helmholtz term, $P^Z - P^H$; whereas the second part, marked as $P^M$ for its meridional predominance, is the remainder of $L$ and the first part. In the semi-implicit Richardson scheme, the first part is taken at the $\mu + 1$ iteration while $P^M$ is lagged behind. This results in a tridiagonal problem

$$[I - \eta(P^Z - P^H)]q_{\mu+1} = q_\mu + \eta [P^M q_\mu - r_{\nu+1}],$$

where $I$ denotes the identity operator, and $\eta$ can be interpreted as a pseudo-timestep, determined from linear stability theory for the $P^M$ operator. For further reference and illustration, we implemented four preconditioning options in the semi-implicit shallow-water model. Options 0 to 2 are elementary, in that: option 0 corresponds to $P = I$, viz. no preconditioning; option 1 is for $P = D$, where $D$ symbolises the diagonal part of $L$ in (10); option 2 is a diagonally preconditioned explicit Richardson iteration for full $L$; and option 3 is the semi-implicit Richardson scheme in (14). Taking the number of GCR iterations with option 0 as a reference, the subsequent options reduce the solver’s iteration count by about 10, 40 and 98 percent, respectively. In the experiments reported in this paper, we use solely option 3 with two Richardson iterations that comprise two tridiagonal inversions and one evaluation of $P^M$. We found that this combination yields the best overall performance.

### 3 Selected test-cases and experimental setup

#### 3.1 Preamble

To assess the impact of mixed precision on the performance of semi-implicit W&C models, we exploit two well-established benchmarks for testing and evaluating shallow-water surrogates of dynamical cores for global weather and climate, namely the Rossby-Haurwitz wave with wavenumber four (RHW4) and a zonal flow ($Q_x/H \propto \cos \phi, Q_y = 0$) past the Earth orography. The Rossby-Haurwitz waves are analytic solutions of the nondivergent nonlinear barotropic vorticity equation on the sphere, and as such propagate zonally without change of shape (Haurwitz, 1940). However, Rossby-Haurwitz waves are incompatible (Temam, 2006) with shallow-water equations (SWE), and their SWE simulations lead to unstable solutions, given sufficiently short zonal wavelength and sufficient amplitude (Hoskins, 1973). In particular when simulated with SWE, the form of RHW4 changes a little over 24 days, while twice shorter Rossby-Haurwitz waves break completely within a week (Hoskins, 1973). Because of this marginal stability (Thuburn & Li, 2000), RHW4 is a convenient vehicle to test capability of numerical methods for maintaining subtle nonlinear balance of wave form solutions over an extended integration time. The RHW4 simulation with SWE was included in the suite of tests proposed by Williamson et al. (Williamson et al., 1992), and has become a prominent benchmark in the field.

As the RHW4 problem is particularly smooth, with the analytic solution confined to a few spherical harmonics (Thuburn & Li, 2000), we complement it with a multi-scale variant of the classical problem of a geostrophically balanced zonal flow past a hill centered in mid latitudes. The original problem has been studied by Grose and Hoskins (Grose & Hoskins, 1979). It is characterised by small Rossby and Froude numbers—here $Re = U/Lf$ and $Fr = U/\sqrt{gH}$, with $U$, $L$ and $H$ denoting the characteristic flow speed, horizontal scale of the hill and the mean height of the shallow-water surface, respectively—whereby it is well explained by the linear theory (Grose & Hoskins, 1979). This problem, also proposed by Williamson et al. (Williamson et al., 1992) for evaluating the efficacy of numerical methods for global-scale dynamics, has become a benchmark in the
To quantify the impact of the reduced precision on the solver convergence and solutions quality, we adapt the $L_2$ (Euclidean) and $L_{\infty}$ (infinity) error norms defined in (Williamson et al., 1992) as normalised deviations from a “genuine” solution. For the RHW4 test-case, the genuine solution is the analytic solution $\Psi_0(x - Vt, y)$, where $\Psi_0 = \Psi(x, y)$ represents the initial condition for the velocity and depth fields—given in eqs. (143), (144) and (146) of (Williamson et al., 1992)—and $V$ is the propagation speed of the analytic Rossby-Haurwitz wave (eq. 142 in (Williamson et al., 1992)). As the analytic wave does not satisfy the shallow-water equations (1)-(3), the deviations of numerical solutions from $\Psi_0(x - Vt, y)$ are expected to grow in time. However, the comparison provides a meaningful gauge for assessing the impact of a reduction in precision. In the orographic flow test-case the genuine solution is specified as a geostrophically balanced zonal flow, steady in the absence of the orography. For both benchmarks, the calculations addressing reduced precisions are conducted on the three regular longitude-latitude grids, with respective resolutions $NX \times NY = 128 \times 64, 256 \times 128$ and $512 \times 256$.

3.2 Highlights of RHW4 simulations

Figure 2 shows the analytic initial condition and the numerical results on the 512×256 grid after one and two periods of the analytic RHW4 when the analytic solution exactly reproduces the initial condition. The departures from the analytic result are clearly seen. After one period, the numerical RHW4 is steeper and somewhat retarded compared to the analytic result. After the two periods, the numerical results is less steep but the phase retardation is more pronounced. The lower resolution results (not shown) look similar, which is not surprising as the RHW4 is well resolved (with 32 grid intervals per wavelength) even on the coarsest grid; cf. Figure 2 in (Smolarkiewicz & Margolin, 1998) that shows a coarse result after 5 days. As expected (see figure 3 in (Thuburn & Li, 2000)), the current solution is visibly stable, which however should not be taken for granted, as irregularities of model discretisation can excite instability of the RHW4 already after 5 days; cf. Figs. 11 and 12 in (Szmelter & Smolarkiewicz, 2010). Figure 3 complements figure 2 with the display of relative departures of the numerical solutions from the analytic results.

To appreciate the strenuousness of the illustrated calculations, consider that the celerity of the external mode ($\sqrt{gH_o} = 280$ m/s) is comparable to the speed of sound, while the zonal length interval of the physical grid in the rings of grid points adjacent to the poles is $\delta x = 7675, 1919$ and 480 m for the three considered regular grids. The explicit solutions, like those discussed in (Smolarkiewicz & Margolin, 1998; Szmelter & Smolarkiewicz, 2010), require then temporal intervals $\Delta t \sim 40, 10$ and 2.5 s, respectively, whereas calculations reported here were conducted with $\Delta t = 800, 400$, and 200 s, respectively. The attained linear rather than quadratic reduction of $\Delta t$ with the reciprocal of $NY$ is due to the specificity of the flow decaying towards the poles. Although the calculations employed weak polar absorbers—with the inverse time scale $\alpha$ increasing linearly from zero at the angular distance $\phi \geq 3\pi/64$ away from the poles to $1/(200\Delta t)$ at the poles—essentially the same solution can be obtained without any absorbers. However, even weak absorbers can improve the solver convergence by improving diagonal dominance of the linear operator (11). For example, the presented results using weak polar absorbers $\alpha = 1/(200\Delta t)$ have a 30% reduced Euclidean norm of the residual $\|r_N\|_2$ on average when exiting the GCR algorithm (on average 4 iterations) compared to using no polar absorbers ($\alpha = 0$ everywhere). The calculations illustrated in Figs. 2 and 3, evinced the maximal Courant numbers 118 and 0.31, correspondingly with respect to the celerity of the external mode and the flow speed.
Figure 2. RHW4, instantaneous normalised free-surface perturbations, $(\Phi(x, y, t) - H_o)/H_o$ ($H_o = 8$ km), with imposed corresponding flow vectors. From the top to bottom, respectively, are the initial condition and the numerical results after 7.38 and 14.76 days, corresponding to one and two periods after which the analytic RHW4 propagates by one and two full wavelengths. The perturbation isolines are plotted with the contour interval 0.025 (zero contour lines are not shown), and the reference 100 m/s flow arrow is shown in the lowest right corner.

3.3 Highlights of orographic flow simulations

In analogy to the RHW4 case, figure 4 shows the numerical results on the $512 \times 256$ grid after 14.76 days, likewise simulated with $\Delta t = 200$ s. Pattern-wise, the depa-
Figure 3. RHW4, the normalised free-surface departure from the analytic result, 
$(\Phi(x,y,t) - \Phi_0(x - \mu t, y))/H_o$, with imposed corresponding departures of the flow vectors. 
The top and bottom panels show the results after $t = 7.38$ and $t = 14.76$ days, respectively. The 
contour interval and the reference flow arrow are the same as in figure 2; zero contour lines are 
not shown and the dashed contours correspond to negative values.

Figure 3. RHW4, the normalised free-surface departure from the analytic result, $(\Phi(x,y,t) - \Phi_0(x - \mu t, y))/H_o$, with imposed corresponding departures of the flow vectors. The top and bottom panels show the results after $t = 7.38$ and $t = 14.76$ days, respectively. The contour interval and the reference flow arrow are the same as in figure 2; zero contour lines are not shown and the dashed contours correspond to negative values.

Features from the initial zonal flows are dramatic, especially in terms of energy and direction of local flows. Quantitatively, the free-surface height perturbations (viz. pressure perturbations) are about 3% and 12% in terms of $L_2$ and $L_\infty$ norms; i.e., consistent with real weather (Smolarkiewicz et al., 2019). For reference, the same measures for the RHW4 simulation are of comparable size, with 3% and 7% respectively. In terms of the kinetic energy of the velocity perturbations, the corresponding $L_2$ and $L_\infty$ norms are 153% and 762% (sic), with the latter reflecting both the direction (say easterly or northerly) and the high speed of local flows ($\sim 50$ m/s) emerging in the high latitudes of the northern hemisphere. The respective values for the RHW4 case are much smaller in comparison, with 37% and 35%. The maximal Courant numbers with respect to the celerity of the external mode is 116, which is about the same as in the RHW4 case. The maximal Courant numbers with respect to flow velocity is 0.71, which is more than doubled compared to the RHW4 case. Nonetheless, the overall performance of the elliptic solver was comparable at 4 GCR iterations per time step. The essential difference between the orographic flow simulation and the RHW4 setups is in the importance of the absorbers, its
Figure 4. Normalised free-surface height perturbations \( (H(x, y, t) - H(x, y, 0))/H_0 \) (here \( H = \Phi + H_0 \) and \( H_0 = 8 \) km), with imposed corresponding flow vectors at \( t = 14.76 \) days. The perturbation isolines are plotted with the contour interval 0.00625 (dashed lines are for negative values and zero contour lines are not shown), and the reference 20 m/s flow arrow is shown in the lowest right corner; the contours of the orography \( H_0(x, y) \) span the range 3 to 7003 m with the 1200 m interval.

The inverse time scale here taken as \( \alpha = 1/(2\Delta t) \) consistently with the theoretical considerations (Prusa, 2018).

3.4 Reduced-precision Experiments

The precision reduction is achieved in three distinct steps. In the first step, precision is reduced from double to single precision (23 significand bits) for the entire shallow-water model. The results of these experiments are presented in §4.1. Based on the experience from the tests with the single precision shallow-water model, a mixed-precision shallow-water model is described in §4.2, that behaves similarly to the double precision reference model. As single precision arithmetic is easily available on modern CPUs, these reduced-precision experiments can be performed on standard hardware. In the third step, a mixed-precision model is presented in §4.3 that is additionally using half precision for most parts of the elliptic solver. Half precision is emulated with the reduced-precision emulator (rpe v5 library (Dawson & Düben, 2017)). Using the emulator enables us to explore a precision reduction beyond single precision with model code that was developed for standard computer hardware.

While we emulate the 10 significand bits as in half precision, we keep the number of exponent bits at 11—the double precision standard—which means that the dynamical range of representable numbers remains high. The results should still be representative for the use of half precision as problems with the dynamical range can often be mitigated via rescaling of variables, see for instance (Klöwer et al., 2019), especially since the elliptic problem is linear by design. In section 4.4, the topic of rescaling is again picked up when performance is measured on actual half precision hardware.

The precision levels for the mixed-precision elliptic solver are identified as the minimum precision required for each step of the algorithm shown in figure 1. Because the elliptic solver is part of the timestepping scheme of a non-linear model, it is not obvi-
ous how this could be achieved via analytical means. Thus, we adopt an experimental
approach with the precision being reduced to levels where the elliptic solver still yields
acceptable results. The elliptic solver’s performance is studied for each of the two test-
cases defined earlier. This covers a wide range of different dynamics, from linear responses
up to turbulent regimes with a multiplicity of scales.

We consider three criteria to measure the performance and viability of the ellip-
tic solver when precision is reduced: i) the solver’s convergence rate; ii) the solver ac-
curacy (13); and iii) the deviation of the model solution from the respective “genuine”
solution in the $L_2$ and $L_{\infty}$ error norms. Criterion (i) aims at keeping the solver’s con-
vergence rate unchanged compared to the reference solver. A decrease in the solver’s con-
vergence rate, and thus an increase in the expected solver iterations, would be accept-
able if each iteration is sufficiently cheap when using reduced precision. However, we would
argue that it is still reasonable to aim for approximately the same number of solver it-
erations in the mixed-precision solver, especially if one aims for optimal or near-optimal
speed-ups close to the theoretical limit. Criterion (ii) is straightforward. As soon as the
precision reduction is introduced into parts of the solver, there is a risk that the resid-
ual errors become corrupted with rounding errors. It is therefore important to show that
the elliptic solver is actually capable of converging to the required accuracy. Criterion (iii)
is more subtle. The solver accuracy is chosen based on the additional error in the model
solution of a double precision reference run, defined as the relative error between the ac-
curacy measures with respect to genuine solutions specified in the last paragraph of §3.1.
Given that the residual-error fields of a double precision and a mixed-precision elliptic
solver will be different—e.g., due to a different spectral composition—even if the required
solver accuracy was reached, the pattern and magnitude of the additional discretisation
error might change as well.

4 Results

In this paper, the focus is primarily on the $512 \times 256$ resolution, the highest of the
three resolutions, as it shows the strongest grid convergence at the poles and is thus ex-
pected to be the most challenging test-case for reduced precision.

To evaluate the solver’s convergence rate and the solver accuracy, the discussed resid-
ual error fields are calculated by using definition (12), in double precision arithmetic,
applied to the updated fluid thickness field $\Phi$. We do not use the minimized residual
error field $r$ from step 2 of the GCR algorithm in figure 1 as the values of $r$ may al-
ready be corrupted by rounding errors.

Given these choices, the convergence plots that are an integral part of the follow-
ing discussion are here introduced in detail to avoid ambiguity. In the convergence plots,
the development of the residual error field in the Euclidean norm, as defined in the exit
condition (13), is shown over successive solver iterations. We number the solver iter-
ations as follows: iteration 0 denotes the initial residual error field $\mathcal{L}(\Phi_0) - \mathcal{R}$, iteration 1 to 3 denote the residual error fields $\mathcal{L}(\Phi_{\nu+1}) - \mathcal{R}$, $\nu = 0, 1, 2$ of the first cycle
of the GCR(3) algorithm. Following this convention, iteration 4 to 6 denote the fields
$\mathcal{L}(\Phi_{\nu+1}) - \mathcal{R}$, $\nu = 0, 1, 2$ of the second cycle of the GCR(3) algorithm. To highlight
the residual error field’s evolution over successive solver iterations, values stemming from
different iterations of the same call to the elliptic solver are connected by lines.

4.1 Single Precision Shallow-Water Model

In double precision arithmetic, the solver is converging steadily throughout all solver
iterations, see figures 5 a) and b). The solver requires 4 solver iterations for both test-
cases which is also the minimal iteration number since we enforce for at least one full
cycle of GCR(3). In 3D atmospheric models GCR(3) requires on average between 4 to
Figure 5. Convergence rates of the RHW4 (left) and the orographic flow test-case (right) for three variants of the shallow-water model, a) and b) use double precision arithmetic, c) and d) single precision, e) and f) a mixture of double and single precision. Convergence is shown for 2-day time interval snapshots, and darker colours indicate later simulation time.

Statistics of fluid thickness increments are shown in table 1. For both test-cases, fluid thickness increments drastically reduce in magnitude over successive solver iterations by 5-6 orders of magnitude. For the RHW4, this process happens more quickly, consistent with the slightly higher convergence rate for this test-case.
Given these results, the significance of using $\epsilon = 10^{-5}$ should be discussed. For the orographic flow test-case, the fluid thickness increments can still be up to several centimeters in magnitude until the fourth solver iteration. Errors on a magnitude of centimeters within a single time-step are comparatively large if put into perspective with Table 1 which shows that the $L_2$-norm of the entire solver increment in $\Phi$ is on the order of only 2 meters itself. For the RHW4, the strongly decreased fluid thickness increments after the first solver iteration suggest that it would be sufficient to only perform one single solver iteration, or at most two. However, double precision shallow-water model experiments with an elliptic solver restricted to only one or two solver iterations respectively lead to model crashes due to exponentially growing instabilities, in the case of a single solver iteration even within the first twelve hours of simulation time.

In comparison to the double precision convergence behaviour, single precision solver convergence rates seemingly stall after the first solver iteration, see figures 5 c) and d).

While there are differences in the convergence, concerning criterion (iii)—the deviations from the genuine solution—the solution quality is still good. The differences in the $L_2$ and $L_\infty$ norms compared to the double precision solution are insignificant. In this, the deviations of a model solution are defined at the example of double precision (DP) for the kinetic energy as follows. It is the kinetic energy of the velocity field differences between the double precision solution and its “genuine” solution: $Q_{DP}/\Phi_{DP} - Q_0/\Phi_0$. The largest difference between double and single precision is found for the $L_\infty$-norm of kinetic energy deviations for the orographic flow with a value of 758% for the single precision shallow-water model, a relative change of $5 \cdot 10^{-3}$ compared to the double precision solution.

Further elaborating on this point, the above-discussed kinetic energy deviations as well as their change when going from double to single precision are shown in figures 6 a)-f). The differences shown in figures 6 e) and f) are the absolute values of the differences between figures 6 a) and c) and figures 6 b) and d), respectively. To be more comparable with the figures showing the model solutions itself, here the square-root of kinetic energy is shown. For both test-cases and both precision levels the deviations in kinetic energy can be up to $50 \text{ m/s}$ for both test-cases.

For RHW4 the deviations from the genuine solution are centered around the wave crests and symmetric between both hemispheres. The absolute values of the differences between the double precision perturbations and the single precision perturbations is shown in figure 6 e). The differences are at least two orders of magnitude smaller than the values of the deviations.

| Solver iteration | 1        | 2        | 3        | 4        |
|------------------|----------|----------|----------|----------|
| RHW4             | $0.6m$   | $5 \cdot 10^{-4}m$ | $2 \cdot 10^{-5}m$ | $5 \cdot 10^{-7}m$ |
|                  | $3.5m$   | $2 \cdot 10^{-3}m$ | $3 \cdot 10^{-4}m$ | $6 \cdot 10^{-6}m$ |
| Orographic Flow  | $1.8m$   | $4 \cdot 10^{-2}m$ | $1 \cdot 10^{-3}m$ | $3 \cdot 10^{-5}m$ |
|                  | $48m$    | $1.1m$   | $4 \cdot 10^{-2}m$ | $8 \cdot 10^{-4}m$ |

Table 1. Size of the increments for fluid thickness $\Phi_0$ for the RHW4 and the orographic flow test-case at different iterations of the elliptic solver for the double precision shallow-water model. The upper values provide the average magnitude and the lower values the maximum value throughout the model run.
For the orographic flow, the deviations from the genuine solution show the locations and magnitude of synoptic-scale eddies, as well as regions with high mountains, such as Himalaya, where fluid thickness is small compared to orography and orography gradients are large making the associated flow exhibit strongly non-linear behaviour. The differences, see figure 6 f), are about one order of magnitude smaller than the deviations themselves.

Differences in model solution between double and single precision need to be further put into perspective. Simulations of atmosphere and ocean models suffer from a number of errors that influence results, including imperfect initial or boundary conditions, discretisation errors, or unknown physics for some of the model components. It is therefore important to put the impact of rounding errors into the context of other sources of errors. If the impact is small, a reduction in precision will likely have no measurable impact on prediction quality. This approach has been applied in a number of publications already (Düben et al., 2014, 2017; Hatfield et al., 2018; Düben & Dolaptchiev, 2015). Here, we add a random perturbation (white noise with an amplitude of five centimeters) to the initial conditions of the double precision shallow-water model of which we know that they will be within the level of initial condition uncertainty. It is found that the differences in kinetic energy deviations between the perturbed and unperturbed double precision simulations are comparable to the differences between the simulations in single and double precision. Thus, the change in deviations when going to single precision will probably be insignificant when compared to the uncertainty range typically encountered in W&C prediction.

In summary, although the solver of the single precision model variant behaves differently, the model still produces valid solutions for our test-cases.

To better understand the convergence behaviour of the single precision shallow-water model, first, the potential impact of the absorbers is studied. However it is found that they behave exactly for the double and the single precision shallow-water model. The RHW4 test-case still works even without a polar absorber, while the orographic flow test-case crashes within the first 24 hours of integration time regardless of whether single precision or double precision is used in the shallow-water model.

The convergence behaviour of the single precision model is further investigated by looking at convergence in different locations of the computational domain. For this, the computational domain is split into a pole region (10 zonal bands wide in the vicinity of each pole) and the remaining computational domain which we here refer to as the interior. For the pole region, convergence looks exactly like the overall convergence shown in figures 5 c) and d). For the interior of the domain, the first observation is that there is no increase in the initial residuals \( r_0 \) when compared to the double precision shallow-water model. This shows that the observed increased values of the initial residual \( r_0 \) stem from the pole region exclusively. Concerning convergence in the interior region, three solver iterations are required to reduce the initial residual field \( r_0 \) by 3 orders of magnitude, showing the importance of performing more than one solver iteration. In contrast to the single precision solver, in the double precision shallow-water model, convergence of the first and its subsequent solver iterations is consistent everywhere, regardless of whether it is for the pole region or the interior. These descriptions are valid for both test-cases.

To further understand why the single and double precision solvers show different convergence behaviours, we take a look at the fluid thickness variable \( \Phi \). The fluid thickness variable is stored in single precision arithmetic and its magnitude is on the order of \( 10^4 \) to \( 10^5 \) meters. In comparison, a typical fluid thickness increment at each time-step is on the order of (see again table 1) \( 3 \cdot 10^{-5} \) to \( 5 \cdot 10^{-7} \) meters for the last solver iteration. However, the relative error of single precision truncation is given by the machine epsilon \( \approx 1.19 \cdot 10^{-7} \), indicating that thickness increments smaller than \( 10^{-2} \) to \( 10^{-3} \) meters cannot reliably be represented in the fluid thickness field \( \Phi \).
To study the effect of the increments that are too small to be represented in single precision, a double precision copy of the initial fluid thickness $\Phi_0$ is introduced into the single precision elliptic solver and the residual error fields are calculated from this double precision variable. The double precision copy of $\Phi_0$ is continuously kept updated by the single precision solver’s fluid thickness increments after each solver iteration. The convergence rates derived in this way, see figure 7, look drastically different and much closer to the convergence rates of the double precision variant of the elliptic solver from figures 5 a) and b). The initial residual is still larger when compared to the double precision simulations, but the residual error is now reduced at a comparable rate.

In the following, the solver’s convergence is investigated for each latitude separately. For this analysis, the residual error fields are calculated from a double precision copy of $\Phi$. Figures 8 a)-d) show the convergence rates for each latitude. Here only results for the orographic flow are shown as the RHW4 test-case behaves qualitatively similar. The differences in the $L_2$-norm of the residual values are only visible near the pole. For the other
Figure 7. Convergence rates of the RHW4 (left) and the orographic flow test-case (right) for the shallow-water model using single precision. In contrast to figures 5 c) and d), the residual error fields are however calculated from a double precision copy of the initial fluid thickness $\Phi_0$ that is updated by the solver increments after each solver iteration.

latitudes, the $L_2$-norms of the residual error fields are indistinguishable by eye for single and double precision.

The accumulation of numerical errors near the poles of the single precision model can be replicated in the double precision model variant when the elliptic solver’s initial residual $r_0$ is calculated with a fluid thickness $\Phi_0$ that was truncated to single precision—even if all model calculations are performed in double precision arithmetic. The spatial distribution of the numerical errors is illustrated in figure 9. The relative error in the $L_2$-norm is around $10^{-5}$ for the majority of latitudes for both test-cases, while the relative numerical errors in the $L_2$-norm at the poles can easily be as large as 100%. That a small perturbation on the order of the single precision epsilon results in relative errors of this magnitude after applying $L$, shows how badly conditioned the linear operator $L$ is near the poles. The resulting solver convergence of the truncated elliptic solver is illustrated in figures 10, showing a clear degradation when compared to the double precision elliptic solver from figures 5 a) and b).

The precision of the fluid thickness variable $\Phi_0$ is thus of major importance for the calculation of the initial residual error field $r_0$ as well as the subsequent convergence behaviour of the elliptic solver. For substantiation, additional experiments are run using stronger polar absorbers. For the RHW4, the single precision model variant is run with $\alpha = 1/(2\Delta t)$ but it is found that, although the numerical errors in the $L_2$-norm of the initial residual is reduced by half, solver convergence is not improved.

4.2 Mixed-precision Shallow-Water Model

Here, a mixed-precision model is described that shows a very similar convergence behaviour when compared to the double precision variant while performing as many calculations as possible in single precision.

First, based on our previous analysis, the fluid thickness field $\Phi$ used for the initial residual $r_0$ calculation in algorithmic step 1 of the elliptic solver and the update of the model variable $\Phi$ in step 2 should be kept in double precision. This motivates an approach to keep the prognostic model variables $\Phi$, $Q_x$, and $Q_y$ in double precision while the costly part, the calculation of their respective tendencies within a time-step, are calculated in single precision arithmetic. Updating the fluid thickness field $\Phi_\nu$ directly in
Figure 8. Latitude-wise initial and final residual in the $L_2$-norm for the orographic flow test-case for three variants of the shallow-water model, a) and b) use double precision arithmetic, c) and d) single precision, e) and f) a mixture of double and single precision.

algorithmic step 2 is replaced by accumulating the solver increments to fluid thickness in a single precision variable $\Delta \Phi$ which is then added to the fluid thickness field $\Phi^n$ when the elliptic solver is exited. Second, the initial residual $r_0$ calculation is performed in double precision to ensure that the relative error in $r_0$ is smaller than $\epsilon$. And third, we find that the forces $R_i^n$—the forcing terms going into the second-order MPDATA operator
**Figure 9.** Latitude-wise relative numerical error in the $L_2$-norm of the initial residual $r_0$ when the elliptic solver’s initial residual is calculated with a fluid thickness variable $\Phi_0$ that was truncated to single precision. Results are shown for the RHW4 (left) and the orographic flow test-case (right) in 2-day time interval snapshots; darker colours indicate later simulation time.

**Figure 10.** Convergence rates of the RHW4 (left) and the orographic flow test-case (right) for the double precision shallow-water model variant where, in contrast to figures 5 a) and b), the elliptic solver’s initial residual $r_0$ is calculated with a fluid thickness variable $\Phi_0$ that was truncated to single precision.

The majority of the model is still performed in single precision. This includes the entire second-order MPDATA operator (first term on the rhs of (7)), and the calculation of the advective velocities $v^{n+0.5}$.

These precision choices already define most of the mixed-precision shallow-water model except for the precision choice for the elliptic solver cycle—algorithmic steps 2 to 4—that is left at single precision. The consequence of this choice is that the latitude-wise $L_2$-norms of the final residual error fields show a reduction of the convergence rate near the poles. The error is smaller (about an order of magnitude) when compared to
the results from figure 8 d), but still clearly visible. The reduced solver convergence at the poles is however not causing an increase of the initial residual error fields near the poles. Thus one can conclude that the rest of the mixed-precision shallow-water model time-step is not affected, possibly also due to the presence of the polar absorber.

The elliptic solver convergence of the resulting mixed-precision shallow-water model is shown in figure 5 e)-f) and figure 8 e)-f). The global, as well as latitude-wise, representation of the solver convergence show that much of the behaviour of the double precision model variant is restored. Figure 8 e) shows that there is no accumulation of numerical errors at the poles anymore for the orographic wave test-case (RHW4 not shown but qualitatively the same). The convergence rate at the poles is slightly reduced, see figure 8 f).

4.2.1 Reformulation of the Initial Residual $r_0$ Calculation

Based on figure 9, a straightforward single precision calculation of $r_0$ was shown insufficient. In the previous section, double precision was instead used for this calculation. However, since the calculation of $r_0$ is a computationally expensive operation, there is a strong incentive to explore further precision reduction via other means.

Rather than using higher precision, for some Krylov subspace methods a popular option to manage the accumulation of round-off errors in the residual error field is a complete recalculation of the residual after a set number of solver iterations (Van Der Vorst & Ye, 2000). However, each recalculation adds another application of the linear operator $L$ to the total cost of the solver. In our setup we only encounter up to 4 solver iterations until solver convergence, which renders a complete recalculation of the residual values to be computationally inefficient. Additionally, if the numerical errors in $r_0$ become as large as the value itself near the poles, recalculation of $r_0$ becomes meaningless.

Here, a different direction is explored. The numerical errors in $r_0$ ultimately result from the truncation of the fluid thickness field $\Phi$ to single precision. It is presented how splitting parts of the residual calculation into the form of a base part plus a correction term can significantly reduce the numerical error in $r_0$. Since the operator $L$ is linear by design reformulating the calculation of $r_0$ can be done with minimal effort.

As a preparation for splitting the operator calculation, first the fluid thickness variable $\Phi$ is split into two parts. Instead of a double precision variable $\Phi$, the fluid thickness field is now chosen to be described by two single precision fields: $\Phi \approx \hat{\Phi} + \tilde{\Phi}$, where $\hat{\Phi}$ denotes a correction to some to-be-defined base fluid thickness field $\Phi_0$.

At the beginning of the simulation, $\hat{\Phi}$ and $\tilde{\Phi}$ are set via the following procedure:

1. $\hat{\Phi}$ is set to the initial fluid thickness truncated to single precision

$$\hat{\Phi} = \text{trunc}(\Phi^0).$$ (15)

2. $\tilde{\Phi}$ is then defined as the difference calculated in double precision arithmetic between $\Phi$ and $\Phi^0$ truncated to single precision

$$\tilde{\Phi} = \text{trunc}(\Phi^0 - \hat{\Phi}).$$ (16)

The average value of $\tilde{\Phi}$ is then on the order of $10^4 m$ to $10^5 m$, while the average value of $\hat{\Phi}$ is $10^{-3} m$—a difference of 7 to 8 orders of magnitude which is consistent with the machine epsilon of single precision. The relative error of the sum $\hat{\Phi} + \tilde{\Phi}$ when calculated in double precision arithmetic is on the order of $10^{-15}$ which is consistent with the double precision machine epsilon.
During the simulation, the solver increments to fluid thickness are added to $\Phi$ each time-step. Thus, the average values of $\Phi$ increase in time which increases the relative error in $\Phi$ and by definition also of the sum $\Phi + \tilde{\Phi}$. After about 100 time-steps, the average values of $\Phi$ are on the order of $10^1 m$, and the relative error of the sum $\Phi + \tilde{\Phi}$ performed in double precision arithmetic increases to about $10^{-11}$. If this continued further, eventually the relative errors of the sum $\Phi + \tilde{\Phi}$ would reach the single precision machine epsilon.

As a result, in our approach we decide to update $\Phi$ after every 100 time-steps. This is a good compromise between precision and computational overhead. The procedure to update $\Phi$ and $\tilde{\Phi}$ works as follows:

1. Temporarily store the sum
   \[ \Phi^{DP} = \Phi + \tilde{\Phi} \]  
   in a double precision variable $\Phi^{DP}$

2. $\Phi$ is set to the initial fluid thickness truncated to single precision
   \[ \tilde{\Phi} = \text{trunc}(\Phi^{DP}) \]  

3. $\tilde{\Phi}$ is then defined as the difference calculated in double precision arithmetic between $\Phi$ and $\Phi^0$ truncated to single precision
   \[ \tilde{\Phi} = \text{trunc}(\Phi^{DP} - \tilde{\Phi}) \]  

Since this update happens so rarely throughout the model simulation, the computational overhead is negligible.

So far, the split of a double precision variable into two single precision variables has not resulted in any computational gains. The trick to make this approach computationally efficient lies in rewriting the gradient calculation of $\Phi$, i.e. the partial derivatives $\frac{\partial \Phi}{\partial x^J}$, in the linear operator $\mathcal{L}$ of definition (11). The rewritten linear operator reads as follows:

\[ \mathcal{L}(\Phi) := \sum_{I=1}^{M} \frac{\partial}{\partial x^I} \left( \sum_{J=1}^{M} A^{IJ} \left( \frac{\partial \tilde{\Phi}}{\partial x^J} + \frac{\partial \tilde{\Phi}}{\partial x^J} \right) + B^I \Phi^{SP} \right) - C \Phi^{SP} \]  

In this, $\Phi^{SP} = \Phi + \tilde{\Phi}$ is the fluid thickness in single precision. This means, our approach requires a third single precision variable for fluid thickness. The $\frac{\partial \Phi}{\partial x^J}$ only change every 100 time-steps. Thus, they can be either precomputed and stored which would result in two more single precision variables, or if memory is crucial, they could be recomputed every timestep. Additionally, compared to the original definition (11), the operator (20) consists of two more floating point operations per grid-point—the addition of the partial derivatives—but each of these is now performed in single precision arithmetic.

To sum up the required computations for the implemented versions of both operators. Each time-step, the original operator requires 19 double precision floating point operations per grid point and 10 double precision variables. In comparison, the operator (20) consists of 22-24 single precision floating point operations and 12-15 single precision variables, depending on whether $\frac{\partial \Phi}{\partial x^J}$ is stored or recomputed. Additionally, every 100th time-step, there is an additional cost of 4 double precision floating point operations.

Comparing figures 9 and the latitude-wise relative numerical errors when $r_0$ is calculated using the single precision operator (20), see figures 11, the relative numerical error is strongly reduced.

Using the new operator (20) in conjunction with the mixed-precision model of §4.2 reveals that using the new operator throughout the entire computational domain is not
Figure 11. Latitude-wise relative numerical error of the initial residual $r_0$ calculated with the operator (20) with respect to the original operator (11) in double precision arithmetic in the $L_2$-norm for the RHW4 (left) and the orographic flow test-case (right) in 2-day time interval snapshots; darker colours indicate later simulation time.

4.3 Mixed-precision with Half Precision for the Elliptic Solver

In this section the aim is to take down precision to half precision for as many parts of the elliptic solver as possible while still satisfying a reasonably high solver accuracy.

The initial residual error $r_0$ calculation in algorithmic step 1 is not considered for half precision arithmetic as the machine epsilon of half precision is $9.77 \cdot 10^{-4}$ which means it is far larger than our targeted solver accuracy.

The focus is here primarily on the preconditioning and the application of the linear operator (11) in algorithmic steps 1 and 3. This is in part motivated by the fact that these steps are by far the most computationally intensive parts of the elliptic solver. It is found that these parts of the solver can indeed be taken down to half precision for the most part. Two exceptions from this rule however need to be made. The first exception concerns the last operation of the linear operator (11) where $C\Phi$ is subtracted from the previous terms. This calculation is done in single precision because $C\Phi$ and the remaining terms of the linear operator (11) are found to be of similar size and the subtraction of both terms introduces large numerical errors if done in half precision. The second ex-
ception is the area close to the poles. Since the operator (11) is ill-conditioned near the poles, half precision is found to be critical in these regions. Thus, for grid-points within a three zonal-band wide range closest to the North- and Southpole all calculations of the preconditioner and the linear operator are performed in single precision.

In comparison to the preconditioning and the application of the linear operator, algorithmic steps 2 and 4 are computationally cheap in terms of the normalized runtimes, consisting only of global sums and variable updates. The large range in the normalized cost of step 4 is due to the size of the sum changing with $\nu$. Concerning the updates of the residual error field $r_{\nu+1}$ and the accumulation of solver increments to fluid thickness in the variable $\Delta \Phi_{\nu+1}$, both are kept in single precision. Both require precision higher than the machine epsilon of half precision to satisfy the solver accuracy. Concerning the update of $L(p_{\nu+1})$, experiments where the sum is calculated in half precision arithmetic immediately reduces the attainable solver accuracy to 4 orders of magnitude for both test-cases. Additionally all the summands are only available in single precision, which makes single precision the preferred option. However, the sum to obtain $p_{\nu+1}$ can be performed in half precision.

Concerning the global sums for $\beta$ and $\alpha_l$ in algorithmic steps 2 and 4, single precision is the straightforward choice for these calculations. This is simply because all of the involved input variables are single precision and truncating them to half precision first would likely result in a computational overhead. Rounding errors in the global sums were on the order of the single precision epsilon and thus no issue for the studied model resolutions.

Figure 12. Convergence rates of the RHW4 a) and the orographic flow test-case b) for the mixed-precision elliptic solver variant with half precision, and the corresponding latitude-wise initial c) and final d) residual error in the $L_2$-norm for the orographic flow test-case.
The elliptic solver convergence of the resulting mixed-precision shallow-water model is illustrated in figure 12 a)-d). The global, as well as latitude-wise, representation of the solver convergence show that much of the behaviour of the double precision model variant is preserved. But there is a clear degradation of solver convergence for the last solver iteration.

Figure 12 c) shows that there is no accumulation of numerical errors at the poles for the orographic wave test-case. In Figure 12 d), the impact of enforcing higher precision in the vicinity of the poles becomes visible. The final values for the latitude-wise residual error $L_2$-norm near the poles are smaller in comparison to the rest of the computational domain, where the effect of low precision manifests in overall larger values when compared to previous results. The corresponding figures for the RHW4 are not shown but look qualitatively the same.

Figure 13. Square-root of kinetic energy deviations in $\frac{m}{s}$ of the mixed-precision (MP) model from the respective “genuine” model solution at $t = 14.76$ days for the RHW4 in a), and the orographic flow test-case in b). Figures c) and d) show the respective difference in kinetic energy deviations between the double precision and mixed-precision shallow-water model solution in logarithmic scale.

Although the final residual error values are generally larger, the deviations from the genuine solution, which are shown in figures 13 a) and b), are found to be very close to the double precision shallow-water model from figures 6 a) and b). The differences in the deviations compared to the double precision model, see figures 13 c) and d), are at least one order of magnitude smaller than the difference found between single precision and double precision from figures 6 e)-f).

Coming back to §3.1, the two coarser model resolutions $256 \times 128$ and $128 \times 64$ were also run using the mixed-precision shallow-water model described in this section. As expected, it is found that the results are qualitatively the same (not shown here) and the analysis that has been done for the $512 \times 256$ applies as well to the coarser model resolutions. The vicinity near the poles for which higher precision levels are used is cho-
4.4 Expected Computational Performance of the Mixed-Precision Solver

Ultimately, we are interested in increasing the solver’s, and thus the entire model’s, efficiency. For the simulations done in §4.1 and §4.2, time measurements are performed to estimate the computational savings. The single precision shallow-water model of §4.1 has a shorter time-to-solution than the double precision shallow-water model, the model runtime is reduced by 35%. The mixed-precision shallow-water model variant of §4.2 was 30% faster than the double precision variant. For larger problem sizes, these differences between double and single precision performance are generally expected to become even larger. This is because larger problem sizes are typically limited by data transfer between memory, cache, and the processor, which is where low precision variables have a clear advantage as more relevant data can be stored closer to the processor. For instance, for state-of-the-art W&C models single precision simulations are typically reporting a decrease of computing time by 40% in comparison to simulations in double precision (Váňa et al., 2017; Maynard & Walters, 2019).

Concerning memory, the used memory is measured for the three different shallow-water model variants. Memory usage is almost halved when going from double precision to single precision, which was expected since all variable are changed to single precision. For the mixed-precision shallow-water model variant of §4.2, memory was found to be 56% to 60% of the memory used in double precision.

For the mixed-precision elliptic solver with half-precision arithmetic (§4.3), the use of half precision is confined to the application of the preconditioner and the linear operator $L$. To study the impact of half precision on the performance of algorithmic step 3, the entire model code is ported to the Fujitsu A64FX, which supports half precision arithmetic from Fortran and is available on the Isambard 2 supercomputer. The model is compiled using the Fujitsu Fortran compiler 4.3.1 with the -kfz option. With the -kfz option, subnormal numbers are flushed to zero, whose treatment on this type of chip would otherwise severely impact performance. Concerning variable rescaling, even though there is significant anisotropy in the linear operator $L$, after rescaling the coefficients are found to fit into the dynamical range of half precision of about 10 orders of magnitude. It was confirmed that the mixed precision variant on the A64FX provides the same results as the emulated reduced precision experiments in §4.3 in terms of the three criteria: convergence rate, solver accuracy and solution quality.

The measured runtimes for the mixed precision variant and the double precision variant of algorithmic step 3 are shown in figure 14 a). These numbers include the rescaling of variables and the type-casting between different precision levels. Time measurements are performed for a wide range of problem sizes until memory limits are reached. The mixed precision variant is consistently faster. Also, it uses only 27% of the memory that is required for double precision. Speed-ups, see figure 14 b), for larger grid sizes lie between 3.8 and 4.2 consistently. For more information on half precision performance on the A64FX for physical models, see (Klöwer et al., 2022).

To keep the porting to Isambard 2 simple, estimates for the total reduction in solver runtime will be based on the savings for algorithmic step 3. Given the measured speed-ups for algorithmic step 3, it is justified to assume a speed-up of factor of 4 for half precision and a respective factor 2 for a reduction to single precision level. The same holds for memory. These assumptions are in accordance with the peak performance of modern processors and are also justified by the increase of cache, memory and bandwidth efficiency. To estimate the total reduction of runtime, we multiply the number of floating point operations that has been performed with the three different precision levels (double, single and half) with their respective estimated cost (1.0, 0.5, 0.25) for the differ-
Figure 14. Measured runtimes a) and Speed-up b) for algorithmic step 3 for the mixed-precision variant (MP) compared to double precision (DP). The measured runtimes include necessary rescaling and changes in precision levels.

ent algorithmic steps. Including the exceptions for precision levels that we made in §4.3, the overall runtime can be expected to reduce by a factor of 3.3, when comparing the mixed-precision elliptic solver with half-precision to the double precision variant.

5 Conclusion

Mixed-precision in a representative semi-implicit shallow-water model in general and its preconditioned elliptic solver in particular is found to be possible and overall advantageous when compared to the default double precision option. The described mixed-precision shallow-water model using double and single precision clearly outperforms the double precision model in terms of time-to-solution while maintaining the same solution quality and similar solver convergence behaviour. Concerning the elliptic solver, most of the computationally expensive components of the solver can be performed almost entirely in half precision arithmetic.

This being said, it is found that the special structure of our problem does not permit a naive reduction of precision. We should expect the exacerbation of artificial modes in the prognostic variable fields of the model when going to too low precision levels in the elliptic solver. In our configuration, precision reduction close to the poles—the two grid-singularities that were present—caused spurious behavior even for the single precision shallow-water model. Overall, as the solver precision is further and further reduced to its limits throughout this paper from double precision to precision levels as low as half precision, a very clear progression is seen in terms of an increase in numerical errors, reduced convergence rates, and the increasing need for assigning higher precision levels for parts of the computational domain.

On the other hand, it is shown that given some knowledge about the specific structure of the application, issues with reduced precision can in fact be mitigated by using higher precision arithmetic in parts of the computational domain. The path towards a computationally efficient mixed-precision elliptic solver can be a challenging one though, as it requires a holistic way of thinking about the interplay between numerical errors from a local precision reduction, to the resulting global model errors, and the solver’s overall convergence rate. Understanding this interplay may sometimes require in-depth analysis and in some cases even the reformulation of parts of the existing solver algorithm. The approach we found might also be beneficial in other applications.
All in all, results are similar to (Baboulin et al., 2009; Carson & Higham, 2018; Haider et al., 2018; Anzt et al., 2019; Göddeke et al., 2007; Idomura et al., 2018; ?, ?; Furuichi et al., 2011) in that the preconditioning step is very robust against reducing precision and seems to be the part of the elliptic solver where performance gains can be expected while many of the other parts can be quite sensitive to reducing precision. It is an important finding that this general structure can also be found in our specific problem from geophysical fluid dynamics, using a conjugated residual solver with a preconditioner based on tridiagonal inversion, a key component in 3D atmospheric semi-implicit grid-point models.

A recurrent theme in previous publications on reduced precision in W&C models is the motivation that there are irreducible uncertainties (such as initial condition uncertainties, stochastic representations of unresolved subgrid-scale processes) present in these models which gives a justification for accepting a certain amount of additional model errors, in the form of reducing precision, in order to increase the overall model’s computational efficiency (Düben et al., 2014; Chantry et al., 2019; Saffin et al., 2020; Düben et al., 2014; Palmer, 2012; Hatfield et al., 2018). It should be noted that in the elliptic solver this idea of accepting a certain amount of additional model error to increase computational efficiency is at the very core of the approach and is inherently represented and used in the process of setting the solver accuracy.

In summary, given our results, we do not see any major obstacle that would prevent us from using reduced precision arithmetic for an elliptic solver applied to a full 3-dimensional atmospheric model. There is much experience within the atmospheric modeling community concerning the behavior of the height/pressure solve via iterative methods on various computational grids which can be used to circumvent possible pitfalls when reducing precision and instead guide the path towards efficient mixed-precision elliptic solvers for W&C models.

Appendix A Coefficients of the Linear Operator

The derived coefficients of the linear operator (11) are presented. In this, the coefficients are indexed by gridpoints that are denoted by $I = 1 \ldots N$ in longitudinal direction and $J = 1 \ldots M$ in meridional direction.

The $A$-components:

$$A^{11}(I, J) = -g(\Phi^a(I, J) - H_0(I, J)) \ MGI_{h_2}(I, J) \ A(I, J) \ C(I, J),$$
$$A^{12}(I, J) = -g(\Phi^a(I, J) - H_0(I, J)) \ MGI_{h_y}(I, J) \ B(I, J) \ C(I, J),$$
$$A^{21}(I, J) = -g(\Phi^a(I, J) - H_0(I, J)) \ MGI_{h_2}(I, J) \ B(I, J) \ D(I, J),$$
$$A^{22}(I, J) = -g(\Phi^a(I, J) - H_0(I, J)) \ MGI_{h_y}(I, J) \ A(I, J) \ D(I, J),$$

the $B$-components:

$$B^1(I, J) = -E1(I, J) \ A(I, J) \ C(I, J) - E2(I, J) \ B(I, J) \ C(I, J),$$
$$B^2(I, J) = -E2(I, J) \ A(I, J) \ D(I, J) + E1(I, J) \ B(I, J) \ D(I, J),$$

and the $C$-component:

$$C(I, J) = 1,$$

where:

$$MGI_{h_2}(I, J) = -\frac{Gh_1}{h_2(I, J)}, \ MGI_{h_y}(I, J) = -\frac{Gh_2}{h_y(I, J)},$$
$$A(I, J) = \frac{1}{1 + GMM(I, J)^2}.$$
\[ B(I, J) = A(I, J) GMM(I, J) \]
\[ C(I, J) = 0.5 \, GH_1 \, h_y(I, J) \]
\[ D(I, J) = 0.5 \, GH_2 \, h_x(I, J) \]
\[ E_1(I, J) = -\frac{GH_1 \, g(\Phi^n(I + 1, J) - \Phi^n(I - 1, J))}{h_x(I, J)} \]
\[ E_2(I, J) = -\frac{GH_2 \, g(\Phi^n(I, J + 1) - \Phi^n(I, J - 1))}{h_y(I, J)} \]

with:

\[ GMM(I, J) = 0.5 \Delta t f(J) \]
\[ h_x(I, J) = a \cos(\phi(J)), h_y(I, J) = a \]
\[ GH_1 = 0.5 \frac{\Delta t}{\Delta x}, GH_2 = 0.5 \frac{\Delta t}{\Delta y} \]

Acknowledgments

Jan Ackmann was funded via the European Research Council project ITHACA (grant No. 741112). Peter D. Dueben gratefully acknowledges funding from the Royal Society for his University Research Fellowship as well as funding from the ESIWACE 2 project. ESIWACE 2 has received funding from the European Union’s Horizon 2020 research and innovation program under grant agreement 823988. NCAR is sponsored by the National Science Foundation. This work used the Isambard UK National Tier-2 HPC Service operated by GW4 and the UK Met Office, and funded by the Engineering and Physical Sciences Research Council EPSRC. No primary Earth system data was used for this research.

References

Adcroft, A., Hill, C., Campin, J.-M., Marshall, J., & Heimbach, P. (2004). Overview of the formulation and numerics of the mit gcm. In Proceedings of the ecmwf seminar series on numerical methods, recent developments in numerical methods for atmosphere and ocean modelling (pp. 139–149).

Amritkar, A., & Tafti, D. (2016). Computational fluid dynamics computations using a preconditioned krylov solver on graphical processing units. *Journal of Fluids Engineering, 138*(1).

Anzt, H., Flegar, G., Grützmacher, T., & Quintana-Ortí, E. S. (2019). Toward a modular precision ecosystem for high-performance computing. *The International Journal of High Performance Computing Applications, 33*(6), 1069–1078.

Baboulin, M., Buttari, A., Dongarra, J., Kurzak, J., Langou, J., Langou, J., ... Tomov, S. (2009). Accelerating scientific computations with mixed precision algorithms. *Computer Physics Communications, 180*(12), 2526–2533.

Carson, E., & Higham, N. J. (2018). Accelerating the solution of linear systems by iterative refinement in three precisions. *SIAM Journal on Scientific Computing, 40*(2), A817–A847.

Chantr, M., Thornes, T., Palmer, T., & Düben, P. (2019). Scale-selective precision for weather and climate forecasting. *Monthly Weather Review, 147*(2), 645–655.

Chen, G., Chacón, L., & Barnes, D. C. (2012). An efficient mixed-precision, hybrid cpu–gpu implementation of a nonlinearly implicit one-dimensional particle-in-cell algorithm. *Journal of Computational Physics, 231*(16), 5374–5388.

Dawson, A., & Düben, P. D. (2017). rpe v5: an emulator for reduced floating-point precision in large numerical simulations. *Geoscientific Model Development, 10*(6), 2221–2230.
Düben, P. D., & Dolaptchiev, S. I. (2015). Rounding errors may be beneficial for simulations of atmospheric flow: results from the forced 1d burgers equation. *Theoretical and Computational Fluid Dynamics*, 29(4), 311–328.

Düben, P. D., McNamara, H., & Palmer, T. N. (2014). The use of imprecise processing to improve accuracy in weather & climate prediction. *Journal of Computational Physics*, 271, 2–18.

Düben, P. D., & Palmer, T. (2014). Benchmark tests for numerical weather forecasts on inexact hardware. *Monthly Weather Review*, 142(10), 3809–3829.

Düben, P. D., Subramanian, A., Dawson, A., & Palmer, T. (2017). A study of reduced numerical precision to make superparameterization more competitive using a hardware emulator in the openifs model. *Journal of Advances in Modeling Earth Systems*, 9(1), 566–584.

Eisenstat, S. C., Elman, H. C., & Schultz, M. H. (1983). Variational iterative methods for nonsymmetric systems of linear equations. *SIAM Journal on Numerical Analysis*, 20(2), 345–357.

Furuichi, M., May, D. A., & Tackley, P. J. (2011). Development of a stokes flow solver robust to large viscosity jumps using a schur complement approach with mixed precision arithmetic. *Journal of Computational Physics*, 230(24), 8835–8851.

Gillard, M., & Benacchio, T. (2022). Ft-gcr: a fault-tolerant generalized conjugate residual elliptic solver. *Journal of Computational Physics*, 110997.

Gödeke, D., Strzodka, R., & Turek, S. (2007). Performance and accuracy of hardware-oriented native-, emulated-and mixed-precision solvers in fem simulations. *International Journal of Parallel, Emergent and Distributed Systems*, 22(4), 221–256.

Grose, W. L., & Hoskins, B. J. (1979). On the influence of orography on large-scale atmospheric flow. *Journal of Atmospheric Sciences*, 2(36), 223–234.

Haidar, A., Abdelfattah, A., Zounon, M., Wu, P., Pranesh, S., Tomov, S., & Dongarra, J. (2018). The design of fast and energy-efficient linear solvers: On the potential of half-precision arithmetic and iterative refinement techniques. In *International conference on computational science* (pp. 586–600).

Hatfield, S., Subramanian, A., Palmer, T., & Düben, P. (2018). Improving weather forecast skill through reduced-precision data assimilation. *Monthly Weather Review*, 146(1), 49–62.

Haurwitz, B. (1940). The motion of atmospheric disturbances on the spherical earth. *J. mar. Res.*, 3(5), 254–267.

Hoskins, B. (1973). Stability of the rossby-haurwitz wave. *Quarterly Journal of the Royal Meteorological Society*, 99(422), 723–745.

Idomura, Y., Ina, T., Yamashita, S., Onodera, N., Yamada, S., & Imamura, T. (2018). Communication avoiding multigrid preconditioned conjugate gradient method for extreme scale multiphase cfd simulations. In *2018 ieee/acm 9th workshop on latest advances in scalable algorithms for large-scale systems (scala)* (pp. 17–24).

Klöwer, M., Düben, P. D., & Palmer, T. N. (2019). Posits as an alternative to floats for weather and climate models. In *Proceedings of the conference for next generation arithmetic 2019* (pp. 1–8).

Klöwer, M., Hatfield, S., Croci, M., Düben, P. D., & Palmer, T. N. (2022). Fluid simulations accelerated with 16 bits: Approaching 4x speedup on a64fx by squeezing shallowwaters. jl into float16. *Journal of Advances in Modeling Earth Systems*, 14(2), e2021MS002684.

Korn, P. (2017). Formulation of an unstructured grid model for global ocean dynamics. *Journal of Computational Physics*, 339, 525–552.

Kühnlein, C., Deconinck, W., Klein, R., Malardel, S., Piotrowski, Z. P., Smolarkiewicz, P. K., ... Wedi, N. P. (2019). Fvn 1.0: a nonhydrostatic finite-volume dynamical core for the ifs. *Geoscientific Model Development*, 12(2),...
Kühnlein, C., & Smolarkiewicz, P. K. (2017). An unstructured-mesh finite-volume mpdata for compressible atmospheric dynamics. *Journal of Computational Physics, 334*, 16–30.

Kurowski, M. J., Wojcik, D. K., Ziemianski, M. Z., Rosa, B., & Piotrowski, Z. P. (2016). Convective-permitting regional weather modeling with cosmo-eulag: Compressible and anelastic solutions for a typical westerly flow over the Alps. *Monthly Weather Review, 144*(5), 1961–1982.

Maynard, C. M., & Walters, D. N. (2019). 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.

Mengaldo, G., Wyszogrodzki, A., Diamantakis, M., Lock, S.-J., Girardo, F. X., & Wedi, N. P. (2019). Current and emerging time-integration strategies in global numerical weather and climate prediction. *Archives of Computational Methods in Engineering, 26*(3), 663–684.

Müller, A., Deconinck, W., Kühnlein, C., Mengaldo, G., Lange, M., Wedi, N., . . . others (2019). The escape project: energy-efficient scalable algorithms for weather prediction at exascale. *Geoscientific Model Development, 12*(10), 4425–4441.

Müller, E. H., & Scheichl, R. (2014). Massively parallel solvers for elliptic partial differential equations in numerical weather and climate prediction. *Quarterly Journal of the Royal Meteorological Society, 140*(685), 2608–2624.

Nakano, M., Yashiro, H., Kodama, C., & Tomita, H. (2018). Single precision in the dynamical core of a nonhydrostatic global atmospheric model: Evaluation using a baroclinic wave test case. *Monthly Weather Review, 146*(2), 409–416.

Palmer, T. (2012). Towards the probabilistic earth-system simulator: A vision for the future of climate and weather prediction. *Quarterly Journal of the Royal Meteorological Society, 138*(665), 841–861.

Prusa, J. M. (2018). Computation at a coordinate singularity. *Journal of Computational Physics, 361*, 331–352.

Prusa, J. M., Smolarkiewicz, P. K., & Wyszogrodzki, A. A. (2001). Simulations of gravity wave induced turbulence using 512 pe cray t3e.

Prusa, J. M., Smolarkiewicz, P. K., & Wyszogrodzki, A. A. (2008). Eulag, a computational model for multiscale flows. *Computers & Fluids, 37*(9), 1193–1207.

Saffin, L., Hatfield, S., Dübne, P., & Palmer, T. (2020). Reduced-precision parametrization: Lessons from an intermediate-complexity atmospheric model. *Quarterly Journal of the Royal Meteorological Society, 146*(729), 1590–1607.

Smolarkiewicz, P. K., Deconinck, W., Hamrud, M., Kühnlein, C., Mozdzynski, G., Szmelter, J., & Wedi, N. P. (2016). A finite-volume module for simulating global all-scale atmospheric flows. *Journal of Computational Physics, 314*, 287–304.

Smolarkiewicz, P. K., Grubišić, V., & Margolin, L. G. (1997). On forward-in-time differencing for fluids: Stopping criteria for iterative solutions of anelastic pressure equations. *Monthly weather review, 125*(4), 647–654.

Smolarkiewicz, P. K., Kühllein, C., & Grabowski, W. W. (2017). A finite-volume module for cloud-resolving simulations of global atmospheric flows. *Journal of Computational Physics, 341*, 208–229.

Smolarkiewicz, P. K., Kühllein, C., & Wedi, N. P. (2019). Semi-implicit integrations of perturbation equations for all-scale atmospheric dynamics. *Journal of Computational Physics, 376*, 145–159.

Smolarkiewicz, P. K., & Margolin, L. G. (1998). Mpdata: A finite-difference solver for geophysical flows. *Journal of Computational Physics, 140*(2), 459–480.

Smolarkiewicz, P. K., & Margolin, L. G. (2000). Variational methods for elliptic problems in fluid models. *Proc. ECMWF Workshop on Developments in Numerical Methods for Very High Resolution Global Models, 5-7 June 2000.*
Smolarkiewicz, P. K., Margolin, L. G., & Wyszogrodzki, A. A. (2001). A class of nonhydrostatic global models. *Journal of the atmospheric sciences, 58*(4), 349–364.

Smolarkiewicz, P. K., & Szmelter, J. (2005). Mpdata: An edge-based unstructured-grid formulation. *Journal of Computational Physics, 206*(2), 624–649.

Smolarkiewicz, P. K., & Szmelter, J. (2011). A nonhydrostatic unstructured-mesh soundproof model for simulation of internal gravity waves. *Acta Geophysica, 59*(6), 1109–1134.

Szmelter, J., & Smolarkiewicz, P. K. (2010). An edge-based unstructured mesh discretisation in geospherical framework. *Journal of Computational Physics, 229*(13), 4980–4995.

Temam, R. (2006). Suitable initial conditions. *Journal of Computational Physics, 218*(2), 443–450.

Thuburn, J., & Li, Y. (2000). Numerical simulations of rossby–haurwitz waves. *Tellus A, 52*(2), 181–189.

Tintó Prims, O., Acosta, M. C., Moore, A. M., Castrillo, M., Serradell, K., Cortés, A., & Doblas-Reyes, F. J. (2019). How to use mixed precision in ocean models: exploring a potential reduction of numerical precision in nemo 4.0 and roms. *Geoscientific Model Development, 12*(7), 3135–3148.

Váňa, F., Düben, P., Lang, S., Palmer, T., Leutbecher, M., Salmond, D., & Carver, G. (2017). Single precision in weather forecasting models: An evaluation with the ifs. *Monthly Weather Review, 145*(2), 495–502.

Van Der Vorst, H. A., & Ye, Q. (2000). Residual replacement strategies for krylov subspace iterative methods for the convergence of true residuals. *SIAM Journal on Scientific Computing, 22*(3), 835–852.

Wang, Q., Danilov, S., Sidorenko, D., Timmermann, R., Wekerle, C., Wang, X., . . . Schröter, J. (2014). The finite element sea ice-ocean model (fesom) v. 1.4: formulation of an ocean general circulation model. *Geoscientific Model Development, 7*(2), 663–693.

Williamson, D. L., Drake, J. B., Hack, J. J., Jakob, R., & Swarztrauber, P. N. (1992). A standard test set for numerical approximations to the shallow water equations in spherical geometry. *Journal of computational physics, 102*(1), 211–224.

Yang, C., Xue, W., Fu, H., You, H., Wang, X., Ao, Y., . . . others (2016). 10m-core scalable fully-implicit solver for nonhydrostatic atmospheric dynamics. In *Sc’16: Proceedings of the international conference for high performance computing, networking, storage and analysis* (pp. 57–68).