

RESEARCH ARTICLE

10.1002/2016MS000862

Key Points:

- To reduce computing cost for superparameterized atmosphere models, a reduction in precision is studied
- A detailed precision analysis is performed that relates numerical precision to model uncertainty
- Results of the precision analysis are used to improve the model for both high and low-precision runs

Correspondence to:

P. D. Düben,
peter.dueben@physics.ox.ac.uk

Citation:

Düben, P. D., A. Subramanian, A. Dawson, and T. N. Palmer (2017), A study of reduced numerical precision to make superparameterization more competitive using a hardware emulator in the OpenIFS model, *J. Adv. Model. Earth Syst.*, 9, 566–584, doi:10.1002/2016MS000862.

Received 14 NOV 2016

Accepted 3 FEB 2017

Accepted article online 11 FEB 2017

Published online 28 FEB 2017

A study of reduced numerical precision to make superparameterization more competitive using a hardware emulator in the OpenIFS model

Peter D. Düben¹, Aneesh Subramanian¹ , Andrew Dawson¹ , and T. N. Palmer¹

¹AOPP, Department of Physics, University of Oxford, Oxford, UK

Abstract The use of reduced numerical precision to reduce computing costs for the cloud resolving model of superparameterized simulations of the atmosphere is investigated. An approach to identify the optimal level of precision for many different model components is presented, and a detailed analysis of precision is performed. This is nontrivial for a complex model that shows chaotic behavior such as the cloud resolving model in this paper. It is shown not only that numerical precision can be reduced significantly but also that the results of the reduced precision analysis provide valuable information for the quantification of model uncertainty for individual model components. The precision analysis is also used to identify model parts that are of less importance thus enabling a reduction of model complexity. It is shown that the precision analysis can be used to improve model efficiency for both simulations in double precision and in reduced precision. Model simulations are performed with a superparameterized single-column model version of the OpenIFS model that is forced by observational data sets. A software emulator was used to mimic the use of reduced precision floating point arithmetic in simulations.

Plain Language Summary Weather and climate models cannot represent physical processes of the Earth System explicitly that are smaller than the distance between model grid points. Due to limitations in computing power, this distance is typically larger than 10 km in simulations of the global atmosphere. However, the spatial scale for many important physical processes, such as clouds, is much smaller than this and large errors are generated for predictions of both weather and climate due to limited resolution. To approximate the behavior of subgrid-scale processes within atmosphere models, superparameterization was developed that is running a two-dimensional small-scale model within each grid column of the global model. Superparameterization can improve model simulations but it causes a very large increase in computational cost in comparison to standard simulations. To reduce computational cost, this paper investigates whether it is possible to reduce numerical precision when running the small-scale model. It is shown that precision can indeed be reduced such that computing costs can potentially be reduced significantly. It is also shown that results of an investigation of reduced numerical precision provide valuable information for the quantification of model uncertainty and model development.

1. Introduction

Most of today's global atmosphere models cannot run at a resolution higher than 10 km for weather and 100 km for climate simulations due to limitations in computing power, and difficulties enabling models to scale efficiently onto large fractions of peta-scale supercomputers. However, limited resolution is a large source of model error since processes that cannot be represented explicitly need to be parameterized within model simulations. In particular, the explicit representation of individual clouds and convection cells is still not possible for typical weather and climate models. This is believed to have a strong negative impact on the quality of model simulations. One of multiple approaches that have been made to improve the representation of subgrid-scale processes in global model simulations is the use of so-called superparameterization schemes [see for example, Grabowski, 2004]. Here a cloud resolving model (CRM) is introduced into every grid column of a general circulation model (GCM). The CRM typically has only two spatial dimensions (one horizontal and the other vertical). The horizontal domain size is typically smaller compared to the grid spacing of the GCM and uses periodic boundary conditions within each GCM grid cell. Hence, the

© 2017. The Authors.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

computing cost is significantly reduced in comparison to three-dimensional simulations with the CRM or a global cloud resolving GCM simulation.

The CRM uses the field values of the individual cells of the global model simulation as boundary condition and feeds tendencies of water vapor and temperature back to the GCM. While superparameterized models have shown that the representation of tropical dynamics can be improved significantly (in particular, the Madden-Julian Oscillation [Kim *et al.*, 2009; Benedict *et al.*, 2014], the diurnal cycle [Kooperman *et al.*, 2014; Pritchard *et al.*, 2011], and the Indian Monsoon [Goswami *et al.*, 2015; DeMott *et al.*, 2011]), the use of the CRM increases computational cost of a superparameterized simulation by a large factor (approximately a factor of 20 in the model that was used for this study) in comparison to standard GCM simulations. This makes it difficult to argue that the increase in computing cost can be justified by improvements in predictions, in particular if superparameterized simulations are compared to simulations with the standard GCM at higher horizontal resolution and improved regional forecast skill that generate approximately the same computational cost.

Reducing numerical precision in high-performance computing (HPC) applications can lead to improved performance, and reduced power consumption. This has been exploited for example when the use of mixed precision has been studied in the literature. Here single-precision arithmetic (32 bits) in model parts that are expensive and do not require high precision is combined with double precision arithmetic (64 bits) in those parts that rely on high numerical precision [see for example, Göddeke *et al.*, 2007; Buttari *et al.*, 2008; Baboulin *et al.*, 2009]. A small number of papers has also considered the use of numerical precision that is lower than single precision, such as half precision arithmetic (16 bits) [for example, Minhas *et al.*, 2014] or the use of flexible precision levels [Tse *et al.*, 2012] to speed-up numerical applications. However, given the great potential for savings, it is surprising that there have not been more studies on the use of low numerical precision in computational fluid dynamics in the past.

If it is possible to reduce numerical precision in Earth System models, savings can be reinvested into the use of more ensemble members or an increase in model resolution or complexity to improve predictions of future weather and climate. In particular, it has been argued that numerical precision should be kept high when the dynamics at large spatial scales are computed while precision should be reduced when dynamics at small spatial scales and notably scales close to the truncation scale are calculated [Palmer, 2012; Thornes *et al.*, 2017]. This is intuitive since the dynamics at small spatial scales are inherently uncertain in atmosphere and ocean models. Diffusion smears out small-scale structures in simulations, small-scale dynamics show fast error doubling times that make predictions difficult and parameterization schemes that can only approximate subgrid-scale processes have a strong impact at these scales. Multiple studies with toy models and spectral dynamical cores have shown that a reduction in numerical precision is indeed possible for large parts of model simulations with no strong impact on model quality and that an approach to reduce numerical precision particularly for small-scale dynamics can promise significant reductions in computing cost [see for example, Düben *et al.*, 2014; Düben and Palmer, 2014].

Given the large increase in computational cost due to the use of superparameterization and the ability to reduce computing cost significantly if numerical precision is reduced with spatial scale, it appears obvious that the use of reduced numerical precision beyond single precision should be investigated for the integration of the CRM of the superparameterized model setup. If a reduction in numerical precision is only done for the CRM, keeping the simulation of the GCM at high numerical precision, this would realize a natural-scale separation within model simulations which is difficult to achieve in standard grid point models. Furthermore, all model fields that feedback from the CRM into the global model simulation will be averaged horizontally over all grid columns of the CRM such that individual rounding errors are likely to average out. If numerical cost for the CRM could be reduced due to the use of reduced numerical precision, this would make superparameterization more competitive compared to GCM simulations at higher resolution.

Hardware that can reduce numerical precision beyond single precision is not available for simulations with a superparameterized model yet. However, the use of half precision arithmetic will be possible soon for simulations with weather and climate models on HPC hardware. This could be realized using NVIDIA GPUs that are based on the Pascal hardware architecture but it is also likely that CPU hardware will support half precision arithmetic in the near-term future due to the demand for fast computing at low precision in the machine learning and deep-learning community (e.g., the Intel's Knights Mill-based Xeon Phi processor).

Specialized processors that can work in IEEE half precision arithmetic do already exist (e.g., ARM processors). However, half precision cannot be used on such hardware using the CRM Fortran code since the model would need to be recoded. This is beyond the scope of this paper. A flexible precision approach that would allow different combinations of numbers of bits in the exponent and significand of floating point numbers beyond the IEEE standard could, for example, be realized on Field Programmable Gate Arrays (FPGAs) [see for example, *Düben et al.*, 2015] or if hardware would be based on the Universal Number (UNUM) standard that suggests to use flexible precision for significand and exponent instead of the IEEE floating point standard [see *Gustafson*, 2015]. However, FPGAs still have to prove that they can be used to compute models that are as complex as weather and climate models and it would involve an intensive coding effort to enable the CRM to run efficiently on FPGAs. UNUM hardware does not yet exist to the best of our knowledge. However, even in the absence of hardware, we believe that tests that investigate the use of flexible precision hardware are useful since they provide benchmarks on how much numerical precision can be saved. In this paper, we use an emulator that is able to mimic the use of half precision and flexible precision within model simulations to study the influence of a precision reduction beyond single precision. However, this approach does not allow expected savings in power to be measured since the emulator will increase, and not decrease, computational cost.

It is known that the CRM produces virtually the same quality in results when double or single precision is used (M. Khairoutdinov, personal communication, 2016). However, if the precision level is reduced strongly beyond single precision for the entire model, it is unlikely that simulations will provide reasonable results and indeed we find that simulation quality degrades quickly when precision is reduced to less than the single-precision standard of 23 bits in the significand for all variables globally. On the other hand, if precision is not reduced globally for all model parameters but instead locally for individual parameters or small blocks of model code we expect that precision can be reduced significantly beyond single precision for most parts of the model, we call this a “fine-grained” approach to reduced precision. The more individual parts of the code with local precision levels (precision areas) there are, the more difficult it becomes to identify the optimal combination of precision levels, since the number of possible combinations of individual precision levels will increase exponentially with the number of precision areas. This presents a significant challenge to performing a fine-grained precision analysis for a complex model with chaotic dynamics such as the CRM, where the response of the model to a reduction in precision may be nonlinear and differ significantly for different weather regimes. A fine-grained approach to a precision reduction will, however, provide useful information on model dynamics and help to improve the model. Here the basic idea is to relate the minimal level of precision that can be used locally, to model error and model uncertainty of these parts of the model, and to use this information to improve the model. Adjusting both hardware and software simultaneously to achieve optimal performance for an application is often called codesign. In this paper, we consider predefined hardware configurations and change the software only. To this end, we use the word redesign instead of codesign for improvements of the model that are motivated by the precision analysis in the rest of the paper. However, the results of this paper may have an influence on future hardware setups. We apply two different approaches to model redesign in this paper: (1) to change the CRM model to allow a stronger reduction in numerical precision and (2) to reduce model complexity by removing parts of the CRM that are found to allow a very strong reduction in numerical precision, indicating that these parts do not have a significant impact on model simulations.

This paper will investigate the impact of a reduction in numerical precision for the integration of the CRM in superparameterized model simulations. For the numerical tests in this paper, we will study a single-column model of the Open Integrated Forecast System (OpenIFS) of the European Centre for Medium-Range Weather Forecast (ECMWF). The CRM that is used in this study has been widely used for many single-column model studies [*Khairoutdinov and Randall*, 2003, and references therein] and was coupled to global models in a Multiscale Modelling framework [*Khairoutdinov et al.*, 2005; *Randall*, 2013].

The main scientific contributions of this paper are:

1. We study the use of reduced precision in a multiscale modeling framework and test whether numerical precision can be reduced significantly beyond single precision (section 3).
2. We suggest an approach to search for the optimal combination of fine-grained precision levels (section 3.1). The approach is automated and relies on very short model simulations such that computational cost of the precision search remains reasonable to allow the application to complex models such as the CRM.

3. We show that a reduced precision analysis provides valuable information to quantify model uncertainty of individual model components and to identify model parts that are of less importance (section 4).
4. We show that a reduced precision analysis can be used to inform model redesign to improve both simulations at double precision and at reduced precision (section 4.3).

Sections 2.1 and 2.2 present the emulator for reduced precision and the superparameterized model. Section 3 explains the search algorithm for the optimal level of precision that is proposed in this paper and presents results for simulations with emulated reduced precision. Section 4 is using the reduced precision analysis of the previous section to quantify model uncertainty and to perform model redesign. Section 5 provides the conclusions.

2. Models and Methods

2.1. The Superparameterization Configuration in the Single-Column Version of OpenIFS

The OpenIFS model is the portable version of the Integrated Forecast System (IFS) at ECMWF. The model is based on the IFS model cycle 40R1. The CRM that is used in this study was developed initially at the Colorado State University [see *Khairoutdinov and Randall*, 2003, and references therein] and has been ported into OpenIFS by Filip Váňa (ECMWF) and Marat Khairoutdinov (Stony Brook University). Although the CRM does not resolve all the turbulent eddy length scales of cloud-scale dynamics, the governing equations do permit such dynamics to be physically represented. The model solves the anelastic equations of motion and is based largely on a large eddy simulation model by *Khairoutdinov and Kogan* [1999]. A detailed description of the model dynamic and thermodynamic equations are discussed in *Khairoutdinov and Randall* [2003]. The model is run at a high horizontal resolution on the order of 1–4 km. It solves for the prognostic equations of liquid water/ice moist static energy, and total nonprecipitating and precipitating water. The CRM explicitly resolves the mesoscale updrafts and downdrafts in convective overturns, but parameterizes the small-scale turbulent eddies that play an important role in mediating the cloud entrainment and mixing with the environment. The subgrid-scale model of the CRM uses a 1.5-order turbulence closure based on prognostic turbulent kinetic energy.

In the standard model the interior CRM model grid columns are arranged in a one-dimensional array in horizontal direction, such that the subgrid model is a two-dimensional representation of the cloud resolving motions. However, the CRM can also be run in a three-dimensional setup. The lateral boundary conditions are periodic with a rigid lid at the top of the model grid that spans all of the troposphere. Newtonian damping is employed in the upper third of the model grid to avoid wave reflections. The boundary layer diffusion and surface flux coupling uses the Monin-Obukhov length-scale similarity.

The model experiments in this study used a horizontal grid spacing of 4 km with 16 adjacent columns. The time step for the model integration is 20 s. The vertical grid spacing increases from 50 m near the surface to 500 m above 5 km. The standard model that is used in subsequent simulations will be two-dimensional with 16 grid points in the horizontal and 108 grid points in the vertical direction ($16 \times 1 \times 108$). The CRM uses large-scale forcing and surface forcing fields either derived from a processed observational data set (such as the Atmospheric Radiation Measurement (ARM) data set [Ackerman *et al.*, 2016]) or from fields of the GCM. The large-scale forcing that is used for the four test cases that are investigated in this study is taken from the four different ARM sites described below. The horizontal and vertical large-scale advective tendencies for temperature and vapor mixing ratio are distributed uniformly at a given level and interpolated linearly in time between the discrete measurements in the large-scale forcing. The domain averaged horizontal winds are nudged toward the observed wind profile with a time scale of 2 h.

The four test cases are described in detail below:

1. Test case 1, TWP-ICE: The Tropical Warm Pool-International Cloud Experiment (TWP-ICE) [May *et al.*, 2008] was held in Darwin. High temporal resolution data of precipitation and the atmospheric state were collected during the experiment with radars and centrally located soundings. These ground-based observations are complimented with reanalysis products that provide a sufficient constraint to construct useful model forcing data sets [Xie *et al.*, 2010]. The TWP-ICE is a widely used data set that describes the state and evolution of the tropical clouds, convection and their interaction with the large-scale flow environment. Data taken during this experiment include several ground-based observations and tens of airborne ascent measurements with cloud Lidar instruments. The experiment spanned across a wide range of convective regimes from shallow boundary layer clouds to deep tropical convection and the scales in

- between. The design of the experiment also included producing a high-quality data set for providing boundary conditions for model simulations as well as detailed validation of convective parameterization. Many convective parameterization development experiments have used the core data set for validation and design. Hence, we also use it as a testbed for our analysis. Start date: 17 January 2006, 0:00 A.M.
2. Test case 2, SGP: The first continental site for the ARM program was the Southern Great Plains (SGP) site in the U. S. [Brown *et al.*, 2002; Sisterson *et al.*, 2016]. This region has a very different regime in terms of convection and climatology compared to the TWP-ICE site. It is a midlatitude and mid-continent site which experiences a very broad range of clouds and atmospheric conditions due to the many different varieties of atmospheric disturbances that pass through this location. This site also experiences a strong diurnal and annual cycle of convection over land which is mainly forced by surface latent and sensible heat fluxes. Hence, it is a good test case to compare our results with those from the Tropics. The observational period displays a shallow boundary layer and a negative surface latent heat flux during the early morning, which then increases during the day. Start date: 15 June 2006, 6:00 P.M.
 3. Test case 3, SCSMEX: The South China Sea Monsoon Experiment (SCSMEX) was a joint atmospheric and oceanic experiment to observe the onset and variability of convection during the summer monsoon period. It was a large international field campaign with instruments deployed from many different countries. The experiment occurred during the summer of 1998 and observations of meteorological and oceanic fields during this period were collected. This experiment provides a very good data set to understand the convective processes during the South China Sea Monsoon period and their interactions with oceanic processes. These observations were more similar in regime to the TWP-ICE Tropical warm ocean convection case, yet they were taken in a very different dynamical regime of the monsoon circulation with strong interaction of convection and dynamics. Start date: 6 May 1998, 6:00 A.M.
 4. Test case 4, KWAJEX: The tropical West Pacific is largely characterized by tropical oceanic cloud populations due to the warm waters over this region. An observational field campaign, named KWAJEX, was designed over the Kwajalein Atoll island as a ground validation for the Tropical Rainfall Measurement Mission (TRMM) satellite program in the late 1990s. The field campaign was designed to observe the physical characteristics of the oceanic convective cloud population over this island in the Republic of the Marshall Islands. The KWAJEX observation strategy was designed to obtain radar observations of precipitation and cloud population in a region complementary to the TRMM-based satellite retrievals of the same fields. Start date: 24 July 1990, 6:00 A.M.

In this paper, the CRM is understood as subgrid-scale parameterization. Therefore, we have a particular interest to achieve high quality for the two model fields of the CRM that feedback into the GCM. These are tendencies of water vapor (q_l) and temperature (t_l). Both quantities are defined at each vertical level and calculated as average over the horizontal dimension of the CRM. All simulations for this paper were run in serial mode as a single processing thread. However, the use of multiple threads in parallel would not alter the results since we assume that bit reproducibility could be achieved for different numbers of parallel threads for all hardware configurations that are considered in this paper.

2.2. An Emulator for Reduced Precision

We emulate the use of flexible floating point precision or half precision within model simulations. The emulator that is used in this paper is an open-source tool developed by two of the authors [Dawson and Düben, 2016a, 2016b]. The emulator replaces all REAL number declarations within the CRM by predefined Fortran TYPES. All floating point operations and floating point assignments that are performed for these TYPES are changed such that the results of operations and assignments are truncated to a prescribed floating point precision.

The emulator can reduce numerical precision of the significand of floating point numbers and it can also mimic the use of IEEE half precision arithmetic with 10 bits to represent the significand and 5 bits to represent the exponent, including the correct representation of denormal numbers. The rounding mode that is used in the emulator is very similar to standard IEEE rounding. However, the emulator version used for this work applies a “round to nearest” rounding mode, IEEE 754 uses a “round to nearest, tie to even” scheme. The difference is a special case when rounding a number where the first bit that is rounded away is “1” and all less significant bits are “0,” which would always be rounded so that a “0” remains in the least significant bit of the result (tie to even).

The emulator allows precision to be set separately for whole blocks of code (e.g., a whole subroutine or program) and also to set the precision independently for individual parameters. If a particular parameter has an assigned precision, this precision will always be used for that parameter regardless of any block-level precision. If parameters of different precisions are used in a floating point operation, the result will be truncated to the precision of the highest-precision input, following the IEEE procedure for mixed precision arithmetic. For intrinsic functions, such as sine and cosine, precision will only be reduced for the result of the function even though the function may actually consist of many floating point operations. A full description of the emulator's behavior is given in Dawson and Düben [2016b].

3. Fine-Grained Reduced-Precision in the CRM

We present a recipe to find the optimal level of precision that can be used for many different model parameters and model fields using an automated search (section 3.1). We focus on two different configurations of reduced precision. The first configuration follows the IEEE standard and reduces precision to half precision (16 bits arithmetic) wherever possible. The second configuration assumes that the used hardware is able to use variable precision for the significand of floating point numbers for individual parameters, and thus can use different precisions for different parts of the model. Results of the precision analysis will be presented in sections 3.2 and 3.3 and the quality of model simulations when precision is reduced will be discussed.

3.1. An Automated Search for Minimal Precision

We study a "fine-grained approach" to reduced precision using different precisions for many model parameters and many parts of the model. However, as the number of individual precision levels that are used increases, more and more simulations will be necessary to identify the optimal precision that should be used for each part of the code. For complex models, such as the CRM or even the entire IFS, these tests generate prohibitive computing costs if the model is integrated for a long time in each precision configuration. However, in a recent paper, we could show for simulations with the Lorenz '95 model on FPGA hardware that short-term simulations over 50 time steps can already provide useful information on the precision level that can be used for long-term simulations [Düben et al., 2015]. We will therefore follow the approach to identify the optimal, fine-grained precision level using cheap short-term simulations. For the precision search, we will consider the precision of each parameter to be independent. This has the advantage that the amount of possible precision combinations that need to be tested reduces drastically.

To make the search for the optimal combination of precision more efficient, we enable the computer to perform tests automatically for all predefined precision levels. However, to do this automated search successfully, we need to define an acceptable level of quality for the computer to decide whether the reduction of model quality due to rounding errors for a specific model is acceptable. To obtain an estimate of the maximal acceptable difference between a simulation in double precision and a simulation in reduced precision, we perform 30 simulations in double precision that use different seeds for the initialization of the CRM. We can then calculate the time mean of the standard deviation for the tendencies of water vapor (q) and temperature (t) for all vertical levels of the CRM individually. The standard deviation will be used as reference for the uncertainty of these parameters. Please note that the tendencies t and q are important for the coupling between the CRM and the GCM since they are the only quantities that feedback from the CRM into the GCM simulation. For all simulations with the CRM in reduced precision, we would like the mean difference to the equivalent simulation in double precision (using the same seed for random perturbations of initial conditions) to be smaller—or at least of similar magnitude—compared to the standard deviations of the control ensemble. We consider the tendencies of q and t for the lower $N = 106$ vertical levels and classify a model simulation with reduced precision as unacceptable when the "error" value E is larger compared to a predefined value α for at least one time step of the GCM model which is run for M time steps. Here E is defined as:

$$E = \max \left(\sum_{i=1}^N \frac{q_i^{rp} - q_i^{dp}}{\sigma_{q,i} \cdot N}, \sum_{i=1}^N \frac{t_i^{rp} - t_i^{dp}}{\sigma_{t,i} \cdot N} \right), \quad (1)$$

with q_i^{rp} , t_i^{rp} , q_i^{dp} , and t_i^{dp} being the respective tendencies of the reduced precision and the double precision simulation at the i th vertical level. $\sigma_{q,i}$ and $\sigma_{t,i}$ are the standard deviation of the control ensemble simulation at the i th level. The maximal error α and the number of time steps M are adjusted to the model using trial-and-error tests. We are using $M = 10$ and $\alpha = 0.1$ for the results in this paper.

In the following automated tests, various precision levels are defined for code units consisting of either individual variables, blocks of code, or entire subroutines. For each code unit under consideration, the model will be restarted several times to identify the optimal number of bits in the significand. In the first model run, the precision level will be set to 0 bit in the significand only for the code unit under consideration while using double precision for everything else. If the model fails to meet the quality requirement outlined above for one of the time steps, the model run is aborted after this time step, the precision level is increased by 1 bit and the model is restarted. This process continues until the model simulation fulfils the quality requirement for all M time steps. When a test was successful, the number of significand bits used for a specific code unit under investigation is stored. This procedure is repeated for all code units. This search is automated by organizing the model runs using run scripts and files that are written and read by the individual model simulations such that the user needs to submit the precision analysis only once to investigate all precision levels. The result of this precision search is a table that provides the number of bits in the significand that should be used for the individual code units that have been tested.

This automated search will provide a guess for the optimal precision levels that should be used. However, if all of the code units are reduced to the precision level that was identified by the automated search, it is very unlikely that this full combination simulation will satisfy the quality control. The errors in different code units will not simply add up, and there will be nonlinear interactions between the different rounding errors such that it is difficult to predict how strong the combined reduction in precision will impact model simulations. It may be that interactions between rounding errors will cause instabilities in the model that lead to runtime failures. Therefore, the quality control needs to be fixed at a very low level to be very sensitive to model errors due to reduced precision.

For simulations that test the use of half precision, a very similar automated search was used. However, the number of bits in the significand will not be increased. Instead, it will be tested only once for each code unit whether the model fulfils the quality control with half precision, or not.

We started the precision analysis only after the first forecast day. The first day is assumed as spin-up phase for the CRM, as the cloud model has to come into balance with the large-scale forcing terms. The emulator for reduced precision uses the exponent size of double precision floating point numbers (11 bits), except when emulating half precision arithmetic. However, it is assumed that the CRM can run in single-precision arithmetic at sufficient model quality. Therefore, the number of bits for single-precision floating point exponents was used when calculating relative savings in the number of bits in the following sections to provide a fair comparison.

3.2. The Half Precision Setup

We use the search algorithm outlined in section 3.1 to find parts of the CRM that can be processed in half precision. The CRM comprises of a large number of variables such that an evaluation of all of these variables individually is beyond the scope of this paper, even with the emulator and the automated search for precision levels. We tested individual precision levels for all floating point numbers and arrays that are defined in the three most important parameter lists of the model that are transferred between subroutines that contain prognostic and diagnostic model fields, the relevant grid information and model parameters such as gravitational acceleration and specific heat. We performed the automated search for precision levels for each of the four test cases. When at least one of the test cases indicates that half precision would not be sufficient for a specific parameter, we set the precision of this parameter to single precision. However, the results of the precision analysis agreed very well between the different test cases. For the 229 parameters and fields that were tested for the three parameter lists, differences in the precision analysis for the four test cases happened for only 20 parameters.

Three fields that were identified to work in half precision by the automated precision search needed to be switched back to single precision to allow simulations that do not show degradations in results. These were one scalar and one vector field that describe the vertical grid spacing for pressure levels, as well as the gas constant for water vapor. However, it was possible to identify these problematic fields by performing only a small number of short simulations. Out of all of the parameters and model fields of the three parameter lists for which the precision level was tested individually, only 31 use single precision in the reduced precision setup. However, it turned out that a lot of the relevant prognostic parameters and their tendencies (e.g., horizontal and vertical velocity, moist static energy, pressure, temperature, and water vapor) need single

Table 1. List of the Subroutines That Have Been Treated With Individual Levels of Precision^a

Subroutine	% of Computing Time	Ratio of Bits That Were Saved for Half Precision (%)	Bits Used in the Significand for Flexible Precision
advect_scalar2d	29.00	0.75	23
precip_fall	8.28	36.2	23
tke_full	5.81	50.0	12
pressure	3.28	50.0	23
esatw_crm	3.00	11.5	19
diffuse_scalar2d	2.94	12.4	23
cloud_diag	2.27	42.5	22
esati_crm	1.39	11.5	22
dtesatw_crm	0.40	45.8	8
dtesati_crm	0.34	29.0	11

^aThe first subroutines that are listed are the most expensive routines in model simulations. The routines esati_crm, dtesatw_crm, and dtesati_crm have been included for consistency since esatw_crm was already considered. The relative cost in the second column was calculated as average of the results of three simulations for each of the four test cases using gprof with -O2 optimization for the GNU Fortran compiler on an Intel Xeon E5630 processor with 48 GB of memory. The third column provides the ratio of bits for local fields that were saved for the half precision setup in comparison to a single precision simulation. The fourth column provides the number of bits that was used to calculate the specific subroutine in flexible precision mode. It should be noted that the relative savings for the half precision simulations do not take into account model fields that are imported from other subroutines or modules. The purpose of the subroutines are listed here: advect_scalar2d, to advect two-dimensional scalar fields; precip_fall, to calculate precipitation fallout; tke_full, to solve the turbulent kinetic energy equation; pressure, to solve for pressure; esatw_crm, to calculate the saturation curve of water vapour; diffuse_scalar2d, to diffuse two-dimensional scalar fields; cloud_diag, to condensate cloud water and cloud ice; esati_crm, to calculate the saturation curve of ice; dtesatw_crm, to calculate the time derivative of the saturation curve of water vapour; dtesati_crm, to calculate the time derivative of the saturation curve of ice.

precision. These fields are defined at every grid point and use up a large amount of storage. Therefore, the total number of bits for the parameters and fields of the three parameter lists was reduced to only 68.58% of the original value in single precision (from 1,597,664 bits in single precision to 1,095,600 bits in the half precision setup). Furthermore, some of the floating point numbers that were allocated are actually not used in the two-dimensional model configuration, such that this ratio will reduce further if these numbers are excluded to allow a fair comparison. However, additional to the savings due to the use of half precision for the parameters of the three parameter lists, we also performed a fine-grained precision analysis for the use of half precision for local floating point numbers in the most expensive subroutines of the model. Table 1 lists the ratio of savings in the number of bits in comparison to single precision.

Figures 1 and 2 show results for relative humidity and cloud fraction in the SCM for the double precision default configuration, the three-dimensional CRM configuration ($16 \times 16 \times 108$) and simulations with two model setups that

use reduced numerical precision. We note that the results from the SGP site test case are qualitatively different from the other test cases, as the nature of convection in the SGP is fundamentally different, with midlatitude meteorology governing the convection and dynamics in this region. The predominant convection in our test case for this site was shallow convection, which can be seen with the evolution of both the high relative humidity as well as the cloud cover in the low levels in this region compared to the deep convective nature in the other test cases. All model parts that were not tested for reduced numerical precision use double precision in the significand (52 bits). Differences between the control simulations in double precision and the three-dimensional simulation are of the same order of magnitude when compared to differences between the half precision simulation and the control. Simulations with the half precision model setup produce results that are perfectly reasonable for all model test cases and no problems are visible.

3.3. The Flexible Precision Setup

Similar to the reduction of precision for half precision, we tested individual precision levels in the significand for all floating point numbers and arrays that are defined in the three most important parameter lists of the model that are transferred between subroutines that contain prognostic and diagnostic model fields, the relevant grid information and model parameters. We chose the highest-precision value that was identified for individual parameters from a precision analysis for each of the four test cases. The results were generally in good agreement for the four test cases. After completion of the automated precision analysis, again three parameters needed to be adjusted to higher-precision levels manually to allow simulations that do not show a strong degradation in results. Again, the same two parameters related to the vertical grid spacing created problems that were also switched manually in the half precision setup. Additionally, the horizontal grid spacing needed to be switched to higher precision. For the fields that were tested for the use of reduced precision, the amount of bits could be reduced to 55.61% compared to the number of bits in single

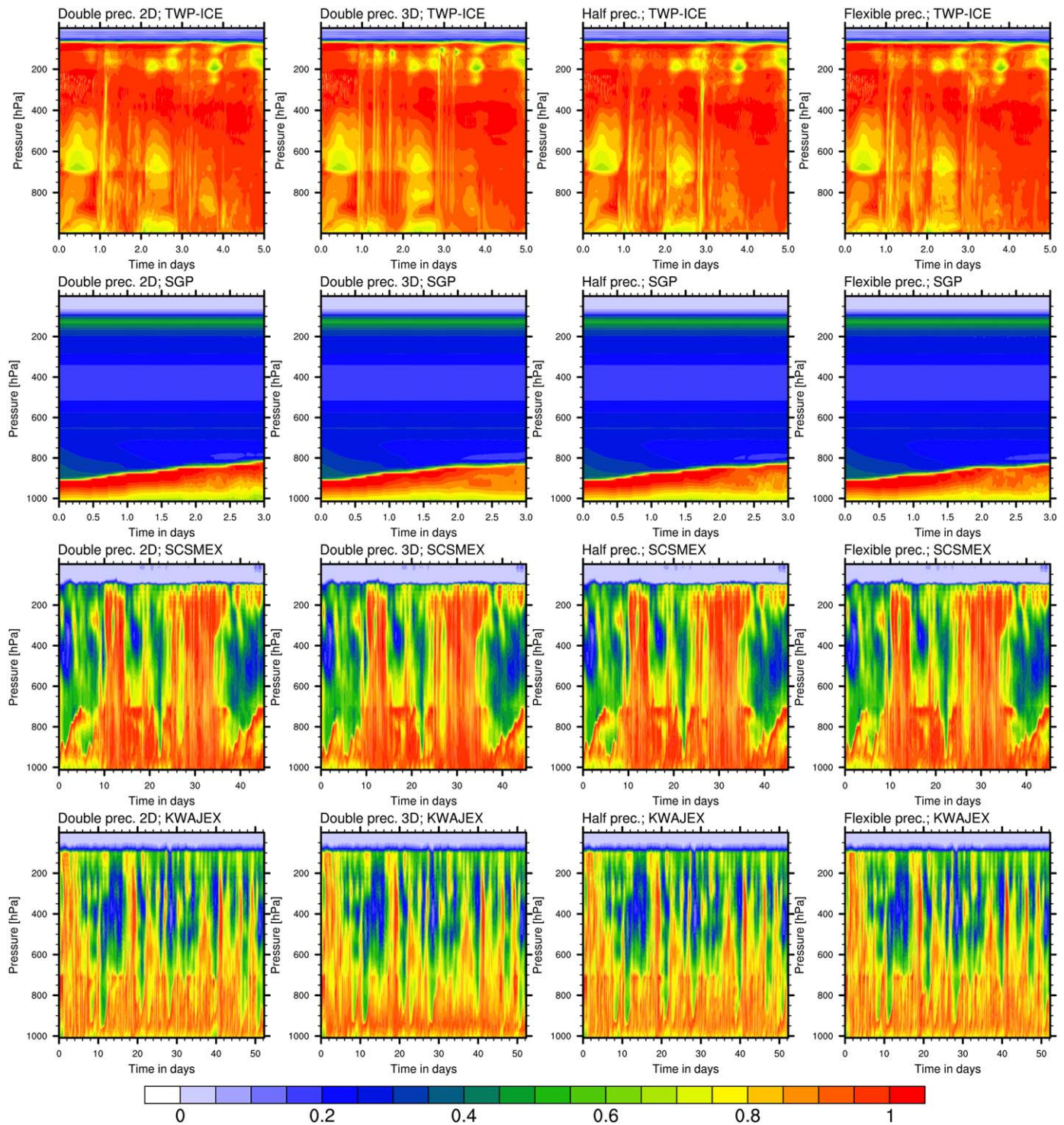


Figure 1. Relative humidity for the reference simulation ($16 \times 1 \times 108$), the three-dimensional CRM simulation ($16 \times 16 \times 108$), the half precision simulation, and the flexible precision simulation (from left to right) for test cases 1–4 (from top to bottom).

precision when all parameters are weighted with their dimension (from 1,597,664 bits in single precision to 888,438 bits in the flexible precision setup). If all parameters that had no impact on the test value ($E = 0.0$ in equation (1) for 0 bit in the significand) were excluded from this consideration, we end up with an average number of 14 bits in the significand, in contrast to 23 bits for single precision.

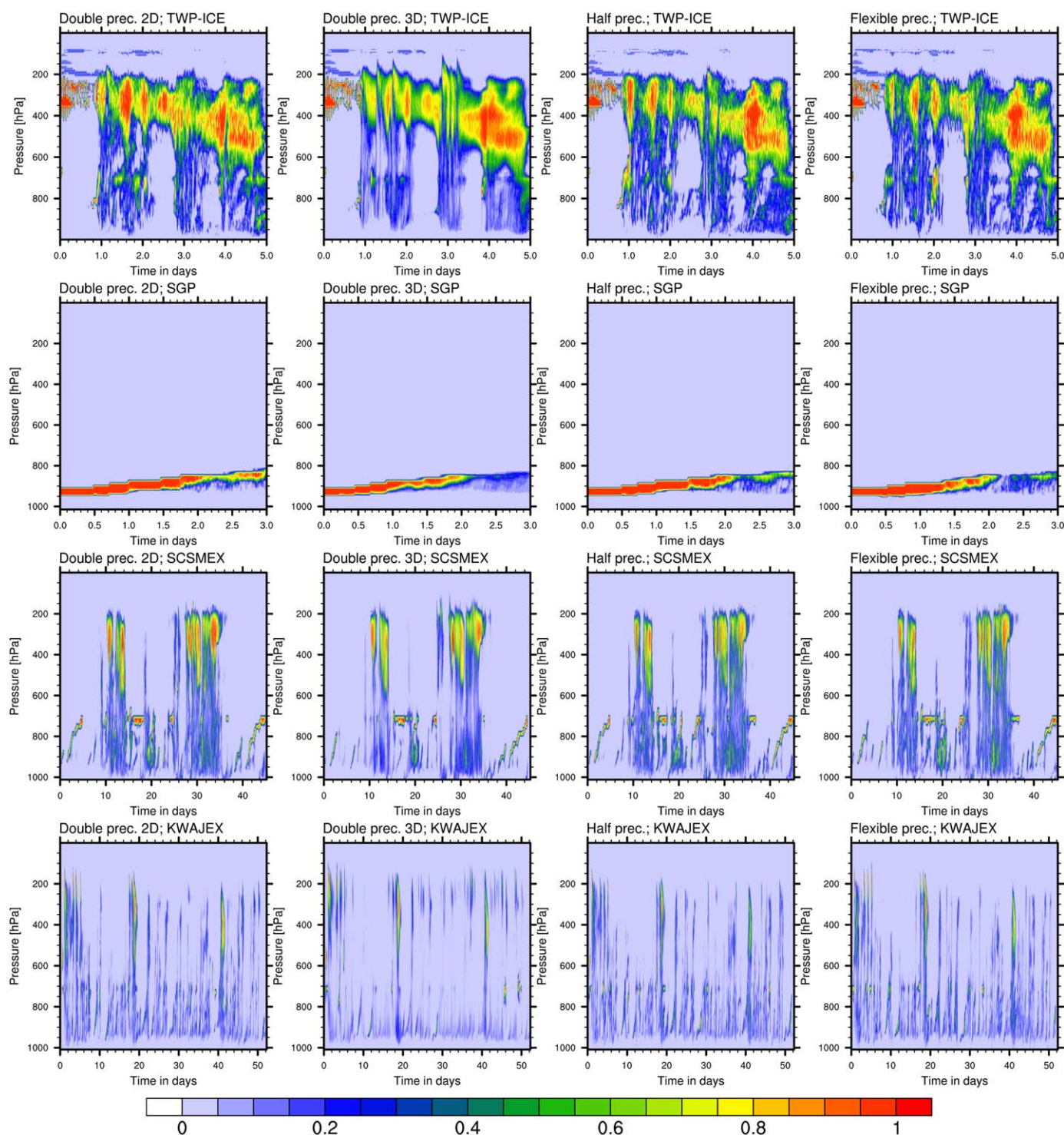


Figure 2. Cloud fraction for the reference simulation ($16 \times 1 \times 108$), the three-dimensional CRM simulation ($16 \times 16 \times 108$), the half precision simulation, and the flexible precision simulation (from left to right) for test cases 1–4 (from top to bottom).

In difference to the tests in half precision, we did not consider individual precision levels for all local variables of the most expensive subroutines in the flexible precision mode. In contrast, we set the precision level of the entire subroutine to a reduced value. Table 1 lists the precision levels that have been identified as optimal. Again, it should be noted that variables that are imported into a subroutine that carry a manually changed individual precision level will not be truncated to the precision level that is defined for the

subroutine. Instead they will use their own precision level within the routine. Furthermore, the precision of values of external floating point fields that are passed into a subroutine that do not have a manually changed individual precision level will not be truncated to the precision level of the subroutine when they enter it. However, these fields will be treated correctly at the precision level of the subroutine in all operations within the subroutine.

Figures 1 and 2 compare results with the flexible precision setup against results with the double precision default configuration and the three-dimensional CRM configuration ($16 \times 16 \times 108$) for relative humidity and cloud fraction in the SCM. All model parts that were not tested for reduced numerical precision use double precision in the significand (52 bits). Differences between the control simulations in double precision and the three-dimensional simulation appear to be of the same order of magnitude when compared to differences between the flexible precision simulation and the control. Simulations with the flexible precision model setup produce results that are perfectly reasonable for all model test cases and no problems are visible.

4. A Reduced Precision Analysis to Quantify Model Uncertainty and to Guide Model Redesign

In this section, we discuss to what extent a fine-grained precision analysis will provide useful information to quantify model uncertainty of model components and to guide model development. In section 4.1, we will present results for ensemble simulations that are based on rounding errors. In section 4.2, we discuss implications of the precision analysis on parameter uncertainty and model uncertainty. In section 4.3, we examine whether the information of the precision analysis can be used to improve the model setup for both high- and low-precision simulations via model redesign.

4.1. Ensemble Forecasts Based on Rounding Errors

It has been argued before, that weather and climate models should represent physical quantities at the level of precision that can be justified by actual information content, and that the use of double precision is overcautious [see for example, *Palmer*, 2014]. It was also argued before that rounding errors can be used to represent subgrid-scale variability and that rounding errors can be used to generate ensemble simulations [*Düben and Dolaptchiev*, 2015]. For the CRM, we know that single precision provides a sufficient level of quality for model simulations. However, it is still not obvious how strong the influence of rounding errors on model simulations will actually be if single precision is used. To answer this question, we will use the emulator for reduced precision to mimic the use of various precision levels between single and double precision.

Figures 3 and 4 compare results for ensemble simulations that are either based on initial value perturbations or on the use of different levels of numerical precision. For ensembles that are based on precision, the only difference between simulations is the precision level that has been used for the significand of floating point numbers which varies between 23 and 52 bits for the thirty ensemble members (single precision uses 23 bits in the significand; for these simulations, the double precision exponent was used). The ensemble spread of the tendencies are comparable for the ensemble simulations based on initial value perturbations for the standard model setup and the ensemble simulations that are based on reduced precision after a bit more than two days of simulations. To this end, rounding errors do have a significant impact on model simulations that is large enough to cause a reasonable ensemble spread even though model quality is not necessarily degraded. Results for the control and the three-dimensional model setup show visible differences between the two simulations. In particular, the standard deviation is smaller for the three-dimensional simulations. This can be explained since the tendencies are averaged over more columns before they are fed back into the GCM in the three-dimensional setup.

4.2. A Precision Analysis to Understand Parameter Uncertainty

The precision analysis of the previous section provides specific levels of precision that should be used to represent model parameters or to calculate specific subroutines. Since a reduction in precision will cause a certain error in the representation of a specific parameter and since it is easy to measure this error, the precision analysis will also provide an estimate for the range in which some of the parameters and subroutines can be changed and stochastically perturbed with no strong impact on model dynamics.

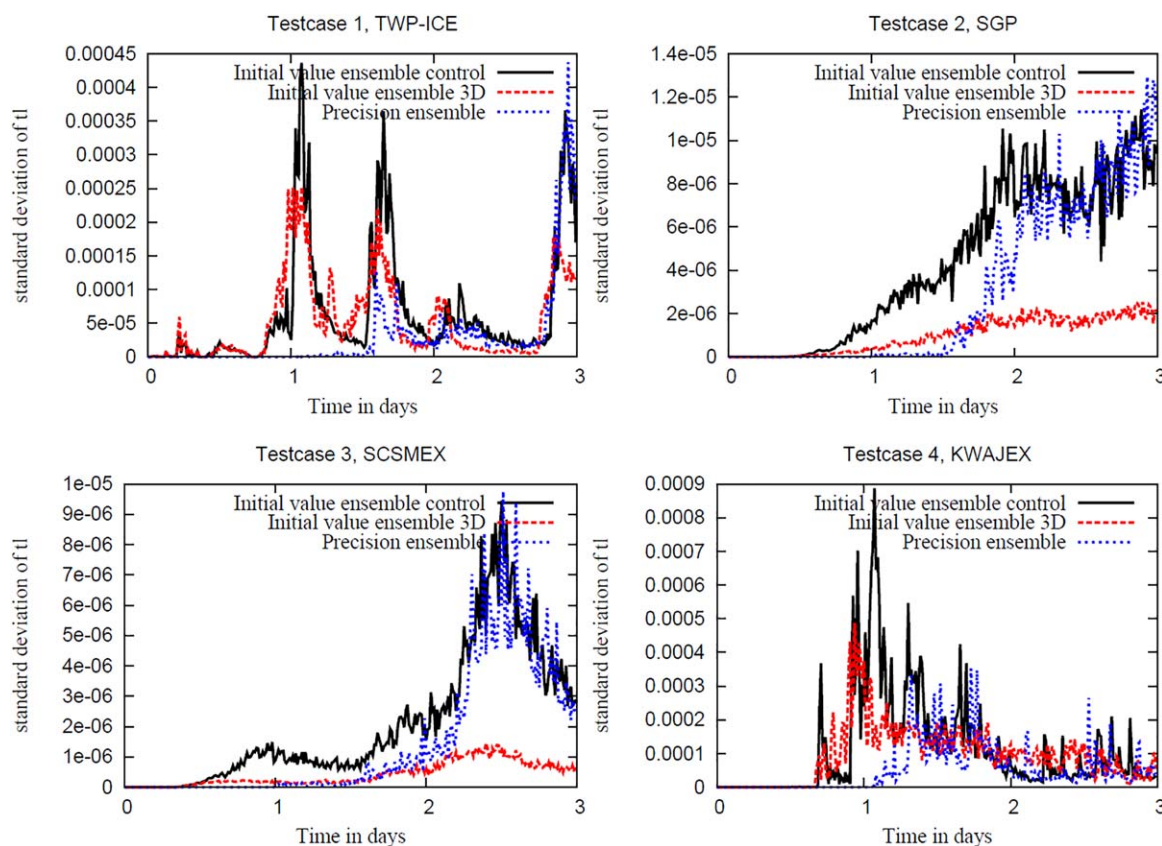


Figure 3. Comparison of the standard deviation of the tendencies of temperature (tl) at the fiftieth vertical level in ensemble simulations with the standard double precision setup and the three-dimensional model setup (both based on initial value perturbations) and ensemble simulations that are solely based on different numbers of bits in the significand for test cases 1–4 (from top left to bottom right).

Tables 2 and 3 provide a list of some of the parameters and variables for which the flexible precision analysis was performed in section 3. Table 2 provides the parameter values that are used in the double precision model setup as well as the values that are used in the reduced precision model setup for the parameters and the relative error. Table 3 provides the precision levels that should be used for the different fields and also the maximal relative rounding error that can happen at this level of precision. Clearly, the representation of parameters as floating point numbers will have an impact on the magnitude of the relative error that a specific parameter will have at a specific precision level. If, for example, a specific parameter would have the value 2.0, the precision analysis would suggest that the parameter can be represented with 0 bit in the significand. However, this would not indicate that the specific parameter could accept a very large level of uncertainty since the representation will still be exact even if precision in the significand is reduced to a minimum. If, on the other hand, a large relative error is indicated by the precision analysis for a specific parameter, this will clearly suggest that the specific parameter carries a large level of uncertainty. For all parameters in Table 2, relative errors due to reduced precision are comparably small (lower than 0.25% for all parameters), which is in close agreement to the certainty of these numbers that are mostly well-known physical parameters. The list of physical fields also appears reasonable with high numerical precision for dynamic variables such as velocity, temperature and pressure and comparably low precision for subgrid-scale fields such as eddy viscosity and eddy conductivity and most of the physical quantities such as condensed liquid water, condensed ice water and precipitating ice water that typically show a large level of uncertainty.

The precision levels that can be used in different subroutines (given in Table 1) also provide valuable information on the impact of these subroutines on overall model dynamics and on uncertainty of the different model parts. It stands out that the turbulent kinetic energy scheme can run with comparably low levels of precision. This indicates that this subroutine does not have a strong impact on model dynamics. We will use this information in section 4.3.1 to optimize the superparameterization model configuration.

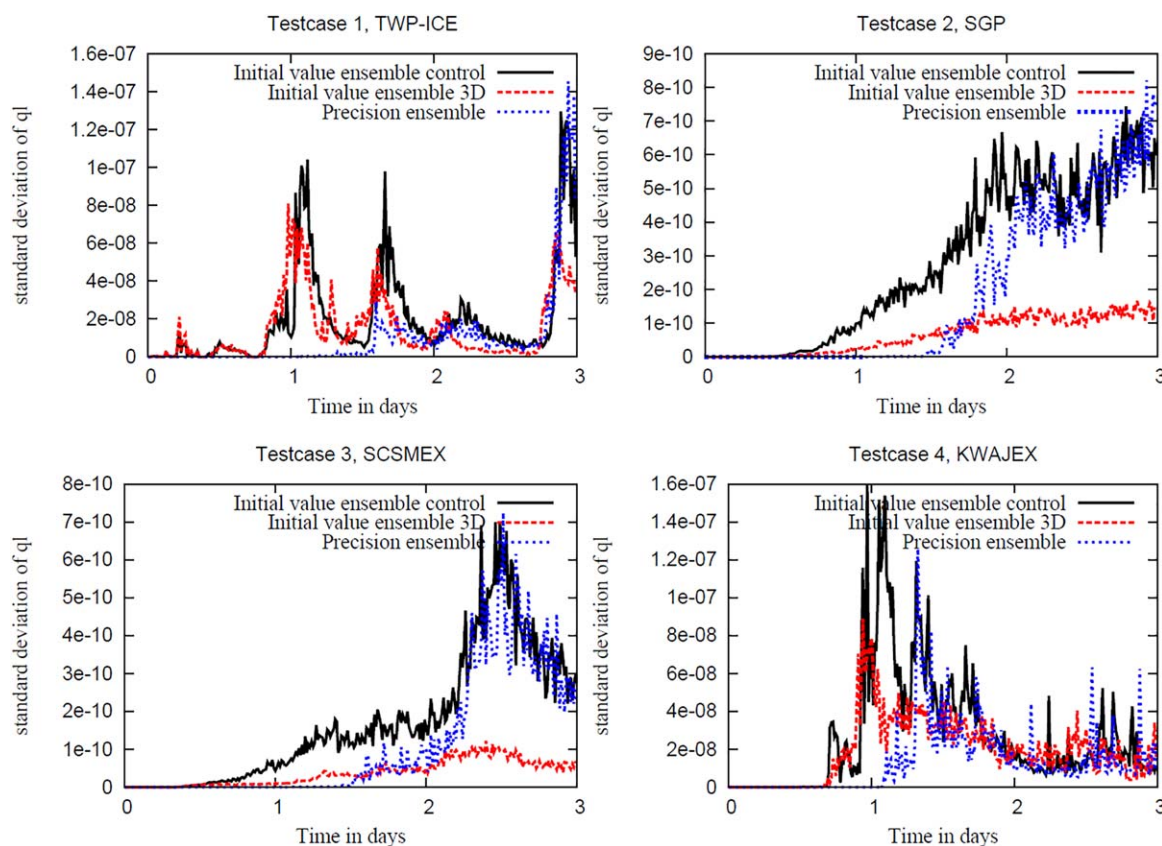


Figure 4. Comparison of the standard deviation of the tendencies of water vapor (q) at the fiftieth vertical level in ensemble simulations with the standard double precision setup and the three-dimensional model setup (both based on initial value perturbations) and ensemble simulations that are solely based on different numbers of bits in the significand for test cases 1–4 (from top left to bottom right).

In future studies, we will investigate whether the uncertainty level that is provided by the reduced precision analysis can be used to improve stochastic parameterization schemes. If the model is, for example, run in single precision it can be assumed that floating point numbers can be randomly perturbed within the range of the rounding errors that were found to be acceptable in the precision analysis with no degradation of model quality. Since the level of uncertainty that is identified by the precision analysis is close to perturbations that will actually influence model results and given the results of the previous subsection that showed a significant spread for ensembles based on comparably high-precision levels, it is likely that a significant ensemble spread can be achieved if the precision levels of the previous section are used to guide the stochastic forcing of parameterization schemes. However, it needs to be explored whether such an ensemble spread accurately quantifies uncertainty in error growth for individual weather regimes.

Table 2. List of Selected Model Parameters and the Precision Level That Was Found by the Automated Precision Search^a

Parameter	Precision	Double Precision	Reduced Precision	Error (%)
Specific heat of air	7	1.004×10^3	1.004×10^3	0.000
Gravitational acceleration	7	9.81	9.8125	0.025
Latent heat of condensation	14	2.5104×10^6	2.510464×10^6	0.003
Latent heat of fusion	7	3.33×10^5	3.33824×10^5	0.067
Latent heat of sublimation	12	2.844×10^6	2.84416×10^6	0.006
Gas constant water vapor	8	4.61×10^2	4.61×10^2	0.000
Diffusivity water vapor	7	2.21×10^{-5}	2.2053719×10^{-5}	0.209
Thermal conductivity of air	8	2.4×10^{-2}	2.3986816×10^{-2}	0.055
Dynamic viscosity of air	3	1.717×10^{-5}	1.7166138×10^{-5}	0.022

^aThe second column provides the number of bits in the significand that was used for the flexible precision setup. The third column provides the value of the parameter in double precision. The forth column provides the equivalent value in the flexible precision representation. The last column provides the relative error of the flexible precision value.

Table 3. List of Selected Model Fields and the Precision Level That Was Found by the Automated Search^a

Model Field	Precision	Maximal Relative Rounding Error
Zonal wind	17	$1/2^{18} \rightarrow 3.81 \times 10^{-4}\%$
Vertical wind	19	$1/2^{20} \rightarrow 9.54 \times 10^{-5}\%$
Moist static energy	23	$1/2^{24} \rightarrow 5.96 \times 10^{-6}\%$
Pressure	22	$1/2^{23} \rightarrow 1.19 \times 10^{-5}\%$
Temperature	23	$1/2^{24} \rightarrow 5.96 \times 10^{-6}\%$
Water vapor	17	$1/2^{18} \rightarrow 3.81 \times 10^{-4}\%$
Condensed liquid water	9	$1/2^{10} \rightarrow 9.77 \times 10^{-2}\%$
Precipitating liquid water	23	$1/2^{24} \rightarrow 5.96 \times 10^{-6}\%$
Condensed ice water	8	$1/2^9 \rightarrow 0.195\%$
Precipitating ice water	8	$1/2^9 \rightarrow 0.195\%$
Subgrid-scale eddy viscosity	3	$1/2^4 \rightarrow 6.25\%$
Subgrid-scale eddy conductivity	6	$1/2^7 \rightarrow 0.781\%$
Air density at pressure levels	14	$1/2^{15} \rightarrow 3.05 \times 10^{-3}\%$

^aThe second column provides information on the number of bits in the significand that was used for the flexible precision setup. The last column provides the maximal relative rounding error that can happen at this precision level.

4.3. Reduced Precision Analysis to Inform Model Redesign

In this section, we will present two approaches that use information of the precision analysis above to improve the model setup. In the first part of the section, we will improve the performance of the standard double precision model setup, by removing model parts that can be integrated with very low levels of numerical precision. In the second part, we will restructure the most expensive subroutine of the model, to allow a stronger reduction in precision.

4.3.1. Redesign to Reduce Cost of the Standard Setup in Double Precision

In this section, we develop a redesigned

version of the CRM, removing two expensive model components for which precision could be reduced significantly. The first change is to remove the turbulent kinetic energy (TKE) scheme from the model. Results of section 3 revealed that the precision of the TKE scheme can be reduced to only 12 bits in the significand with no strong impact on the dynamics of the CRM. For two of the four test cases, the identified precision was actually much smaller (0 and 1 bit in the significand). This indicates that the TKE scheme may not have a strong impact on model behavior. Therefore, we removed the TKE scheme for the redesigned version of the model. If the TKE scheme is removed, two more additional subroutines become obsolete that are solely used by the TKE scheme.

A second change to the model reduces the polynomial order of the saturation curves of water vapor and its derivative from nine to four in the redesigned version of the CRM. It was found in section 3 that the coefficients of the leading polynomial orders can be represented in half precision with no strong impact on model results. However, since these coefficients are very small and actually represented as zero in half precision, this indicates that the polynomial order of the saturation curve can be reduced. Figures 5 and 6 show results of simulations of the control double precision setup (first column) and simulations that use the double precision redesigned setup (second column) that confirm that the reduction of model complexity in the redesigned setup did not cause a significant change in model behavior. The redesigned version reduces model runtime by 12% in comparison to the standard setup. This number was calculated as average of five runs with the default and the redesigned setup for each of the four test cases.

4.3.2. Redesign to Allow a Stronger Reduction in Precision

The CRM was not written with the use of reduced precision in mind. It is therefore likely that precision can be more strongly reduced to a level much closer to the actual information content of specific variables if the model code is changed. The positive definite advection scheme with nonoscillatory option based on Smolarkiewicz and Grabowski [1990] is the most expensive subroutine in the CRM, accounting for more than one fourth of the computational cost (see Table 1). However, the precision analysis suggests that this subroutine does not allow model simulations at high quality if precision is reduced beyond single precision (23 bits in the significand; see Table 1). If we, despite our knowledge from section 3, use half precision for the entire subroutine, model runs produce results that deviate significantly from the control simulation (see third column in Figure 5). However, a closer investigation to identify those places where precision is actually lost within the subroutine reveals that comparatively simple code changes will allow the use of half precision for almost the entire subroutine with no strong degradation in model quality.

Our approach to redesign the model is based on a simple principle. We consider the following differential equation:

$$\frac{\partial \psi}{\partial t} = RHS(\psi(t), p(t)), \quad (2)$$

where ψ is a prognostic field, RHS is the right-hand-side of the equation and p represents other variables of the model. If this differential equation is discretized in time using an explicit Euler method, for the sake of simplicity, we obtain the following equation:

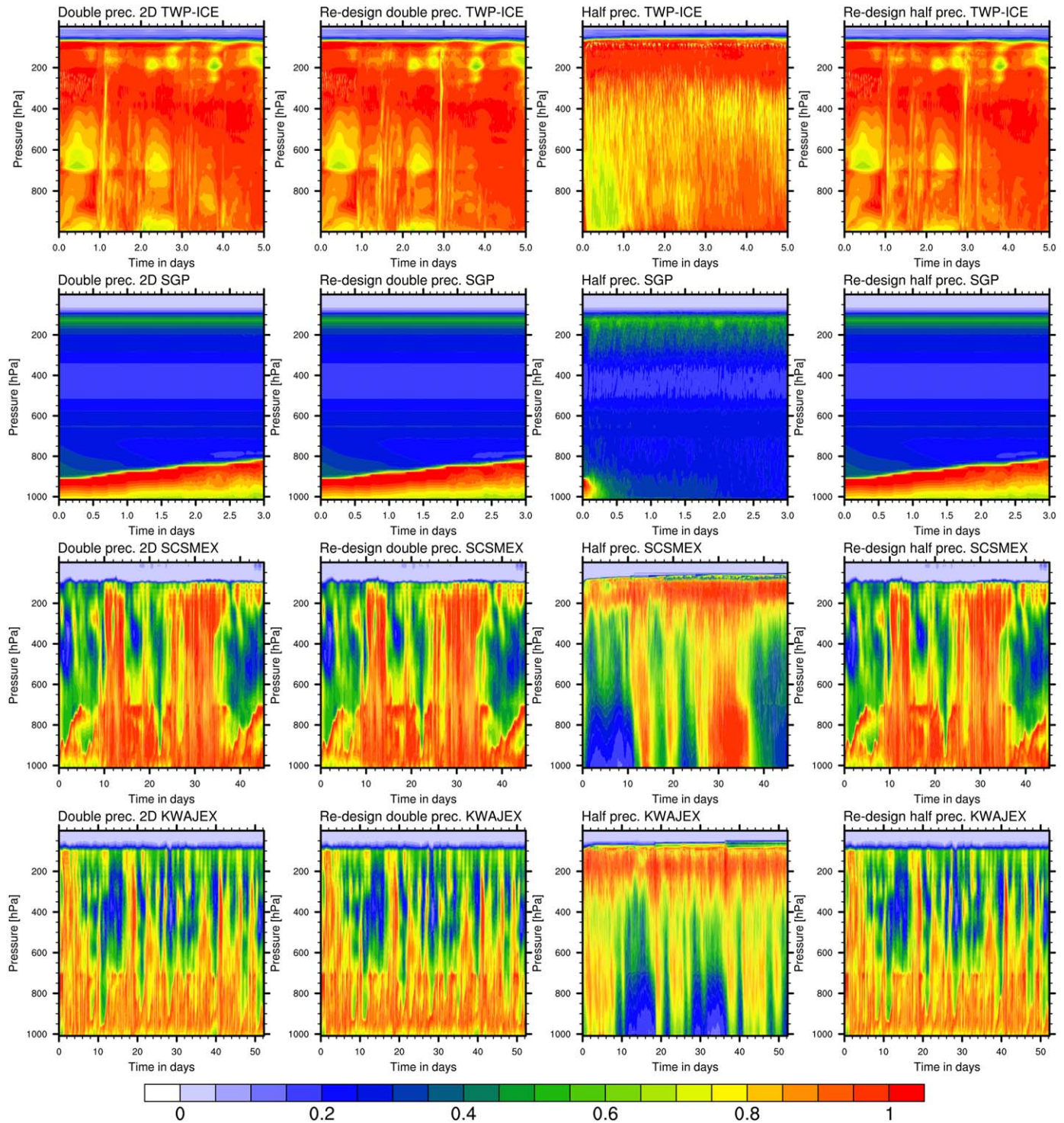


Figure 5. Relative humidity for the double precision control simulation, the redesigned simulations in double precision with no TKE scheme and lower order saturation curves for water vapor to reduce computing cost (section 4.3.1), a simulation with the standard model that is using half precision to calculate scalar advection (section 4.3.2), and a simulation with the redesigned model that is using half precision to calculate scalar advection (from left to right) for test case one to four (from top to bottom).

$$\psi(t^{n+1}) = \psi(t^n) + \Delta t \cdot RHS(\psi(t^n), p(t^n)). \quad (3)$$

To obtain meaningful results the time step is typically small and the increment $\Delta t \cdot RHS(\psi(t^n), p(t^n))$ is typically much smaller compared to the actual field. If precision is reduced, the sum of the field value of the

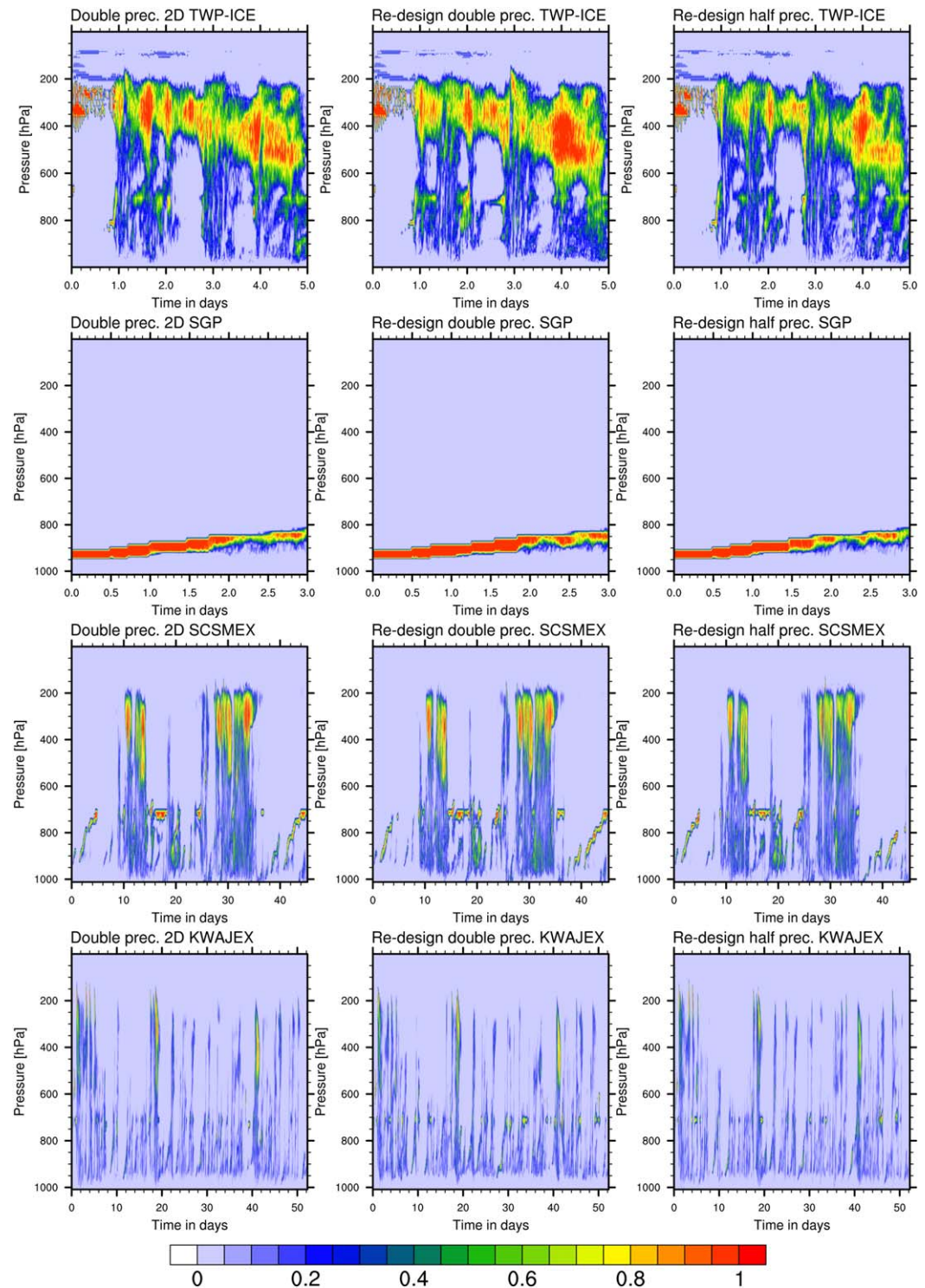


Figure 6. Cloud fraction for the double precision control simulation, the redesigned simulations in double precision with no TKE scheme, and lower order saturation curves for water vapor to reduce computing cost (section 4.3.1) and a simulation with the redesigned model that is using half precision to calculate scalar advection (section 4.3.2; from left to right) for test cases 1–4 (from top to bottom).

previous time step $\psi(t^n)$ with the increment of the time step will shift the floating point exponent of the increment to the same level as the exponent of the physical field, before adding the two significands. If the increment is very small, it will be rounded to zero when the exponent is changed. In this case the variable ψ will not be updated and the differential equation is not represented correctly.

However, if the field ψ is represented at high precision, the increment $\Delta t \cdot \text{RHS}(\psi(t^n), p(t^n))$ can often be represented at a much lower precision level and still allow a good representation of the differential equation. In general, the update can be achieved by first casting the lower precision term manually to higher precision before computing the update, but in practice Fortran compilers will implicitly increase the precision of both summands to the highest precision of the individual summands before performing a sum, which removes the need for explicit casting. If precision for the increment is changed before the sum is performed, the change of the exponent will not impact strongly on the value of the increment when performing the sum. Often the calculation of the increment can be done with copies of the prognostic model fields that use precision values much lower compared to the precision values of the prognostic variables themselves. To reduce precision as strongly as possible, it is therefore often possible to calculate the next time step as follows:

$$\psi(t^{n+1}) = \psi(t^n) + \Delta t' \cdot \text{RHS}(\psi'(t^n), p'(t^n)), \quad (4)$$

where $\psi'(t^n)$, $p'(t^n)$, and $\Delta t'$ are representations of $\psi(t^n)$, $p(t^n)$, and Δt at lower precision. In the redesigned model, the contribution of advection to individual prognostic variables will be calculated with half precision, using a half precision copy of the scalar field to calculate the increment. However, the scalar field itself will still be represented in single precision. This way, the entire subroutine, except for the update of the scalar field, can be calculated in half precision.

Results for simulations that use the redesigned advection scheme in half precision are shown in the fourth column of Figure 5 and the third column of Figure 6. The redesigned simulation and the control simulation in double precision appear to be of the same quality for both quantities. However, computational cost can be assumed to be reduced significantly as almost the entire advection scheme is now using half precision arithmetic.

5. Summary and Conclusions

In this paper, we aimed to reduce computational cost when integrating the cloud resolving model (CRM) of a superparameterized model configuration, to make superparameterization more competitive compared with standard GCM simulations at higher resolution. We have performed a fine-grained precision analysis for many of the fields and parameters of the CRM that is used for superparameterization in a single-column OpenIFS model which is itself driven by observational data. We have used an emulator for reduced precision to test two different modes of reduced precision: the use of half precision for as many fields and parameters as possible and the use of flexible precision in the significand of floating point numbers. We have presented a comparably cheap method to identify the optimal precision for many different parts of the model using an automated precision search. This is a nontrivial task for a model with nonlinear dynamics that is as complex as the CRM. Our method is based on a comparison of the magnitude of changes for short-term forecasts due to a reduction in numerical precision for a specific parameter or model field against the mean spread of ensemble simulations that are based on initial value perturbations. We have performed the automated precision tests for 10 time steps of the GCM for four different model test cases, to study different atmospheric regimes. For both precision setups, we have tested for the optimal precision level of the most important parameters and fields that are shared between subroutines (229 in total) and we also tested for the local use of reduced precision in the most expensive subroutines. After the automated precision search had finished, only three parameters needed to be switched manually to higher-precision levels for each precision setup to perform model simulations with no visible degradation in model quality.

Making estimates for possible savings when using reduced numerical precision is very difficult since we cannot measure savings when using an emulator for reduced precision. Model simulations with the CRM model with reduced precision on real hardware are still not possible. However, recent hardware developments suggest that calculation in half precision will soon be possible (e.g., using Pascal GPUs from NVIDIA or Intel's Knights Mill Xeon Phi processors). The downside of the fine-grained precision approach is that the use of different precision levels in local computations will produce overheads. However, since the processing of floating point numbers is typically considered to be much less resource demanding in comparison to data movement, such overheads can be assumed to remain small.

The results of this paper suggest that savings due to the use of reduced precision could be significant since precision levels can be reduced significantly beyond single precision in large parts of the CRM. However, the results indicate that the use of flexible precision and half precision will achieve an increase in performance in comparison to single precision that is smaller than a factor of two (four when compared to double precision). To this end, reduced precision could make superparameterization more competitive. However, the cost increase due to the use of superparameterization in comparison to a standard model simulation will still be a significant burden that needs to be justified. If the reduced precision approach is combined with running the CRM on GPUs at reduced precision, while the GCM is run on CPUs at high precision on a hybrid hardware architecture, the time latency of the GCM computations having to wait for the CRM computations to finish may be overcome. Furthermore, there is scope that changes of the model configuration may allow significant cost reductions for simulations with the CRM. Jones *et al.* [2015] have recently shown that the use of mean-state acceleration can speed-up model simulations by a factor of 2–8 without seriously degrading accuracy for superparameterized simulations for which the turbulent circulation and clouds evolve faster than the horizontal mean state. If all of the approaches discussed above are combined to speed-up the CRM simulations as much as possible, superparameterization may become significantly cheaper and much more competitive in comparison to standard GCM simulations. An order of magnitude increase in computing speed may be possible.

In the second part of the paper, we have argued that a reduced precision analysis provides valuable information on model uncertainty. We have shown that variable precision can be used to generate ensemble simulations of significant spread when compared to ensembles based on initial value perturbations, even when precision values higher than single precision are used. We have argued that the reduced precision values that are found for model parameters and model fields in the first part of this paper represent a reasonable estimate for the uncertainty of these parameters and fields since the error due to rounding does not cause a degradation of model results while a stronger reduction in precision would cause such a degradation in model quality. The magnitude of rounding errors due to a specific precision level can easily be quantified. We have therefore postulated that the tests toward reduced precision could also be interpreted as tests for model uncertainty and could be repeated with random noise terms instead of rounding errors to improve ensemble spread in ensemble simulations. The precision levels that were identified as optimal could therefore be used to develop stochastic parameterization schemes.

We have shown that a reduced precision analysis provides valuable information for future model development since model parts that do not have a significant influence on model dynamics can be identified since they show a very low level of numerical precision that should be used. To this end, we have presented a redesigned model configuration that removes the TKE scheme and reduces the polynomial order when calculating the water vapor saturation curve that shows no degradation in model quality but reduces computational cost by 13% for double precision simulations. We have also shown that precision can be reduced much stronger in the most expensive subroutine—to calculate scalar advection of two-dimensional fields—when the model formulation is changed.

One possible criticism of the study in this paper could be that all four test cases are used to find the minimal precision level and no independent test case is done to test whether the precision levels can also be used for other weather regimes. However, the test cases span a reasonable amount of different weather phenomena and actually only ten time steps of the GCM are used for precision calibration for each test case (each time step is 900 s long) while results for model simulations are shown for several days for each test case. Therefore, only a very small fraction of the observational data was used for precision tuning. However, we acknowledge that we have not tested all possible scenarios, such as critical point behavior and tipping points. Until model quality is evaluated with reduced precision in global model simulation at full spatial scale, we still need to be careful not to generalize the current results since simulations may drift away from realistic dynamics for climate simulations, e.g., if the TKE scheme is removed in regions with strong boundary layer turbulence. However, these tests are beyond the scope of this paper that focuses on the single column model and will be investigated in a separate study in the future. It would also be interesting to investigate how a change in resolution in the CRM will affect precision levels that should be used. If the precision analysis would only be performed for the three tropical ocean test cases (test cases 1, 3, and 4) ignoring the results of the southern great planes test case 2 that shows a different flow structure, the same reduced precision setup would be found for half precision and only one precision value would be different (lower) for the flexible precision setup.

This paper provides another example that numerical precision can be reduced significantly beyond single precision in large parts of atmosphere models and the results confirm that an approach to reduce numerical

precision with spatial scale in a multiscale modeling framework is a good strategy. The results of this paper show that a fine-grained precision analysis is possible even for complex models that simulate chaotic systems. Even in case numerical precision is eventually not traded against increased performance, a fine-grained precision analysis can help to improve today's models and to understand model uncertainty in different parts of the complex models in use.

Acknowledgments

We thank Marat Khairoutdinov and Filip Vána for their work on the superparameterization configuration of IFS and the single-column model. We also would like to thank Glenn Carver for the OpenIFS support rendered for the superparameterization implementation. The authors received funding from an ERC grant (toward the Prototype Probabilistic Earth-System Model for Climate Prediction, project reference 291406). The data used as initial conditions for simulations can be found in the references that are provided within the paper. No primary data have been generated within this study.

References

- Ackerman, T. P., T. S. Cress, W. R. Ferrell, J. H. Mather, and D. D. Turner (2016), The programmatic maturation of the arm program, *Meteorol. Monogr.*, 57, 3–1.
- Baboulin, M., A. Buttari, J. Dongarra, J. Kurzak, J. Langou, P. Luszczek, and S. Tomov (2009), Accelerating scientific computations with mixed precision algorithms, *Comput. Phys. Commun.*, 180(12), 2526–2533, doi:10.1016/j.cpc.2008.11.005.
- Benedict, J. J., E. D. Maloney, A. H. Sobel, and D. M. W. Frierson (2014), Gross moist stability and MJO simulation skill in three full-physics GCMs, *J. Atmos. Sci.*, 71(9), 3327–3349, doi:10.1175/JAS-D-13-0240.1.
- Brown, A., et al. (2002), Large-eddy simulation of the diurnal cycle of shallow cumulus convection over land, *Q. J. R. Meteorol. Soc.*, 128(582), 1075–1093.
- Buttari, A., J. Dongarra, J. Kurzak, P. Luszczek, and S. Tomov (2008), Using mixed precision for sparse matrix computations to enhance the performance while achieving 64-bit accuracy, *ACM Trans. Math. Software*, 34(4), 17:1–17:22, doi:10.1145/1377596.1377597.
- Dawson, A., and P. D. Düben (2016a), An emulator for reduced floating-point precision in Fortran, doi:10.5281/zenodo.154483. [Available at <https://github.com/aopp-pred/rpe>.]
- Dawson, A., and P. D. Düben (2016b), rpe v5: An emulator for reduced floating-point precision in large numerical simulations, *Geosci. Model Dev. Discuss.*, in review, doi:10.5194/gmd-2016-247.
- DeMott, C. A., C. Stan, D. A. Randall, J. L. Kinter, and M. Khairoutdinov (2011), The Asian monsoon in the superparameterized CCSM and its relationship to tropical wave activity, *J. Clim.*, 24(19), 5134–5156.
- Düben, P. D., and S. I. Dolaptchiev (2015), Rounding errors may be beneficial for simulations of atmospheric flow: Results from the forced 1D Burgers equation, *Theor. Comput. Fluid Dyn.*, 29, 311–328.
- Düben, P. D., and T. N. Palmer (2014), Benchmark tests for numerical forecasts on inexact hardware, *Mon. Weather Rev.*, 142, 3809–3829.
- Düben, P. D., H. McNamara, and T. N. Palmer (2014), The use of imprecise processing to improve accuracy in weather & climate prediction, *J. Comput. Phys.*, 271, 2–18.
- Düben, P. D., F. P. Russell, X. Niu, W. Luk, and T. N. Palmer (2015), On the use of programmable hardware and reduced numerical precision in earth-system modeling, *J. Adv. Model. Earth Syst.*, 7, 1393–1408, doi:10.1002/2015MS000494.
- Göddeke, D., R. Strzodka, and S. Turek (2007), Performance and accuracy of hardware-oriented native-, emulated-and mixed-precision solvers in fem simulations, *Int. J. Parallel Emerg. Distrib. Syst.*, 22(4), 221–256, doi:10.1080/17445760601122076.
- Goswami, B. B., R. P. M. Krishna, P. Mukhopadhyay, M. Khairoutdinov, and B. N. Goswami (2015), Simulation of the Indian summer monsoon in the superparameterized climate forecast system version 2: Preliminary results, *J. Clim.*, 28(22), 8988–9012.
- Grabowski, W. W. (2004), An improved framework for superparameterization, *J. Atmos. Sci.*, 61(15), 1940–1952, doi:10.1175/1520-0469(2004)061<1940:AIFFS>2.0.CO;2.
- Gustafson, J. L. (2015), *The End of Error: Unum Computing*, *Comput. Sci. Ser.*, vol. 24, 2nd corrected printing, 1st ed., CRC Press, Boca Raton, Fla.
- Jones, C. R., C. S. Bretherton, and M. S. Pritchard (2015), Mean-state acceleration of cloud-resolving models and large eddy simulations, *J. Adv. Model. Earth Syst.*, 7(4), 1643–1660, doi:10.1002/2015MS000488.
- Khairoutdinov, M., D. Randall, and C. DeMott (2005), Simulations of the atmospheric general circulation using a cloud-resolving model as a superparameterization of physical processes, *J. Atmos. Sci.*, 62(7), 2136–2154.
- Khairoutdinov, M. F., and Y. L. Kogan (1999), A large eddy simulation model with explicit microphysics: Validation against aircraft observations of a stratocumulus-topped boundary layer, *J. Atmos. Sci.*, 56(13), 2115–2131.
- Khairoutdinov, M. F., and D. A. Randall (2003), Cloud resolving modeling of the arm summer 1997 IOP: Model formulation, results, uncertainties, and sensitivities, *J. Atmos. Sci.*, 60(4), 607–625.
- Kim, D., et al. (2009), Application of MJO simulation diagnostics to climate models, *J. Clim.*, 22(23), 6413–6436.
- Kooperman, G. J., M. S. Pritchard, and R. C. J. Somerville (2014), The response of us summer rainfall to quadrupled CO₂ climate change in conventional and superparameterized versions of the NCAR community atmosphere model, *J. Adv. Model. Earth Syst.*, 6, 859–882, doi:10.1002/2014MS000306.
- May, P. T., J. H. Mather, G. Vaughan, C. Jakob, G. M. McFarquhar, K. N. Bower, and G. G. Mace (2008), The tropical warm pool international cloud experiment, *Bull. Am. Meteorol. Soc.*, 89(5), 629–645.
- Minhas, U. I., S. Bayliss, and G. A. Constantinides (2014), *GPU vs FPGA: A Comparative Analysis for Non-standard Precision*, pp. 298–305, Springer, Cham, Germany.
- Palmer, T. N. (2012), Towards the probabilistic earth-system simulator: A vision for the future of climate and weather prediction, *Q. J. R. Meteorol. Soc.*, 138(665), 841–861.
- Palmer, T. N. (2014), More reliable forecasts with less precise computations: A fast-track route to cloud-resolved weather and climate simulators?, *Philos. Trans. R. Soc. A*, 372(2018), 1–14.
- Pritchard, M. S., M. W. Moncrieff, and R. C. J. Somerville (2011), Orographic propagating precipitation systems over the united states in a global climate model with embedded explicit convection, *J. Atmos. Sci.*, 68(8), 1821–1840.
- Randall, D. A. (2013), Beyond deadlock, *Geophys. Res. Lett.*, 40, 5970–5976, doi:10.1002/2013GL057998.
- Sisterson, D., R. Peppler, T. Cress, P. Lamb, and D. Turner (2016), The arm southern great plains (sgp) site, *Meteorol. Monogr.*, 57, 6–1.
- Smolarkiewicz, P. K., and W. W. Grabowski (1990), The multidimensional positive definite advection transport algorithm: Nonoscillatory option, *J. Comput. Phys.*, 86(2), 355–375.
- Thornes, T., P. D. Düben, and T. N. Palmer (2017), On the use of scale-dependent precision in earth system modelling, *Q. J. R. Meteorol. Soc.*, doi:10.1002/qj.2974, in press.
- Tse, A. H. T., G. C. T. Chow, Q. Jin, D. B. Thomas, and W. Luk (2012), *Optimising Performance of Quadrature Methods with Reduced Precision*, pp. 251–263, Springer, Berlin.
- Xie, S., T. Hume, C. Jakob, S. A. Klein, R. B. McCoy, and M. Zhang (2010), Observed large-scale structures and diabatic heating and drying profiles during twp-ice, *J. Clim.*, 23(1), 57–79.