SPECIAL PROJECT PROGRESS REPORT

Progress Reports should be 2 to 10 pages in length, depending on importance of the project. All the following mandatory information needs to be provided.

Reporting year	July 2015 – June 2016
Project Title:	The use of imprecise arithmetic to increase resolution in atmospheric models
Computer Project Account:	spgbtpia
Principal Investigator:	Prof. Tim Palmer Co-Investigators: Peter Düben, Andrew Dawson, Samuel Hatfield, David MacLeod, Stephen Jeffress, Aneesh Subramanian, Tobias Thornes
Affiliation:	University of Oxford
Name of ECMWF scientists collaborating to the project	Dr Glenn Carver, Dr Martin Leutbecher, Dr Filip Vana, Dr Antje Weisheimer
Start date of the project:	01.01.2014
Expected end date:	31.12.2016

Computer resources allocated/used for the current year and the previous one
(if applicable)
Please answer for all project resources

		Previous year		**Current year**	
		Allocated	Used	Allocated	Used
High Performance Computing Facility	(units)	10,000,000	9,507,057	15,000,000	448,709*
Data storage capacity	(Gbytes)				

*We will perform large simulations with the reduced precision superparametrisation setup in the second half of this year such that we will almost certainly use up all available units until the end of this year.

Summary of project objectives

(10 lines max)

The aim of this project is to study the limits and prospects of the use of inexact hardware in a global atmosphere model (IFS). The use of inexact hardware has the potential to reduce the computational cost significantly due to a reduced energy demand and/or an increase in performance. Furthermore, savings in data storage can be anticipated. If computing cost can be reduced, savings can be reinvested to run models at higher resolution or with an increased number of ensemble members to achieve better predictions for weather and climate. We investigate a reduction of numerical precision in atmospheric applications of different complexity (from idealised test setups to the entire IFS forecast model). In this project, a reduction in precision is realised using either single precision (instead of double precision that is typically used) or the emulation of a stronger reduction of precision within model simulations.

Summary of problems encountered (if any)

(20 lines max)

None.

Summary of results of the current year (from July of previous year to June of current year)

This section should comprise 1 to 8 pages and can be replaced by a short summary plus an existing scientific report on the project

We studied the use of reduced precision arithmetic in several different projects that are summarised in the following.

IFS in single precision

We continued the evaluation of the single precision version of OpenIFS that we have developed in the previous year. Motivated by our study, Filip Vana and other researchers at ECMWF have now also programmed a single precision version of the forecast model of IFS and introduced a switch to choose between single and double precision in the latest IFS cycle. Results with the IFS model confirm our results from the evaluation of OpenIFS and show that both ensemble simulations and long-term simulations in single and double precision are virtually of the same quality for simulations at T399 resolution. However, single precision simulations are up to 40% faster in comparison to double precision simulations on the Cray supercomputer at ECMWF. In a collaborative effort between our working group and scientists at ECMWF, we have written a paper that summarises the results of simulations with IFS in single precision that was submitted to the Monthly Weather Review. The use of single precision appears to be a promising candidate to reduce computational cost of operational forecasts at ECMWF significantly. However, further tests at higher spatial resolution will still be necessary to verify whether simulations at single precision are degraded in comparison to double precision simulations at the resolution of operational weather forecasts.

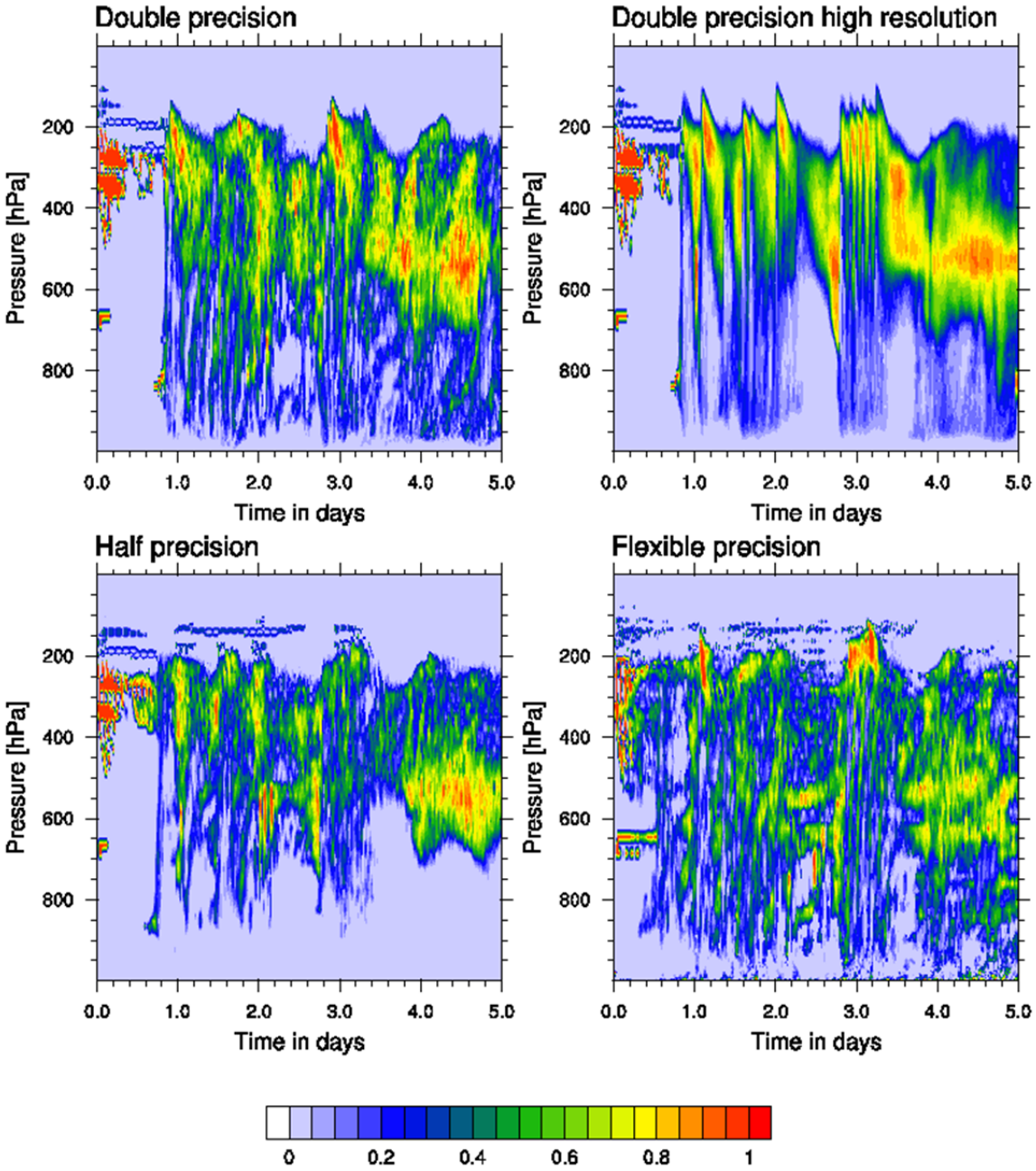
An efficient emulator for reduced numerical precision in simulations of IFS

To study a stronger reduction of precision beyond single precision in simulations of IFS and the global atmosphere, we started to introduce an emulator for reduced numerical precision into parts of OpenIFS and IFS. The emulator was developed in Oxford and is working with type declarations and overloaded operators to allow the emulation of reduced precision in large Fortran models. Only a limited amount of changes in the model code is necessary to enable the use of emulated reduced precision. To this end, real number declarations are replaced by declaration of a predefined type at

the beginning of subroutines and modules. Thereafter, all operations that are performed with theses types are described by the library of the emulator. This allows changes in the precision of floating point operations in both operations and assignments of floating point numbers. The precision level that is used can also be changed locally or even for individual parameters. The emulator is now published as open source on github (https://github.com/aopp-pred/rpe).

A superparametrised version of IFS with reduced numerical precision in the cloud-resolving model

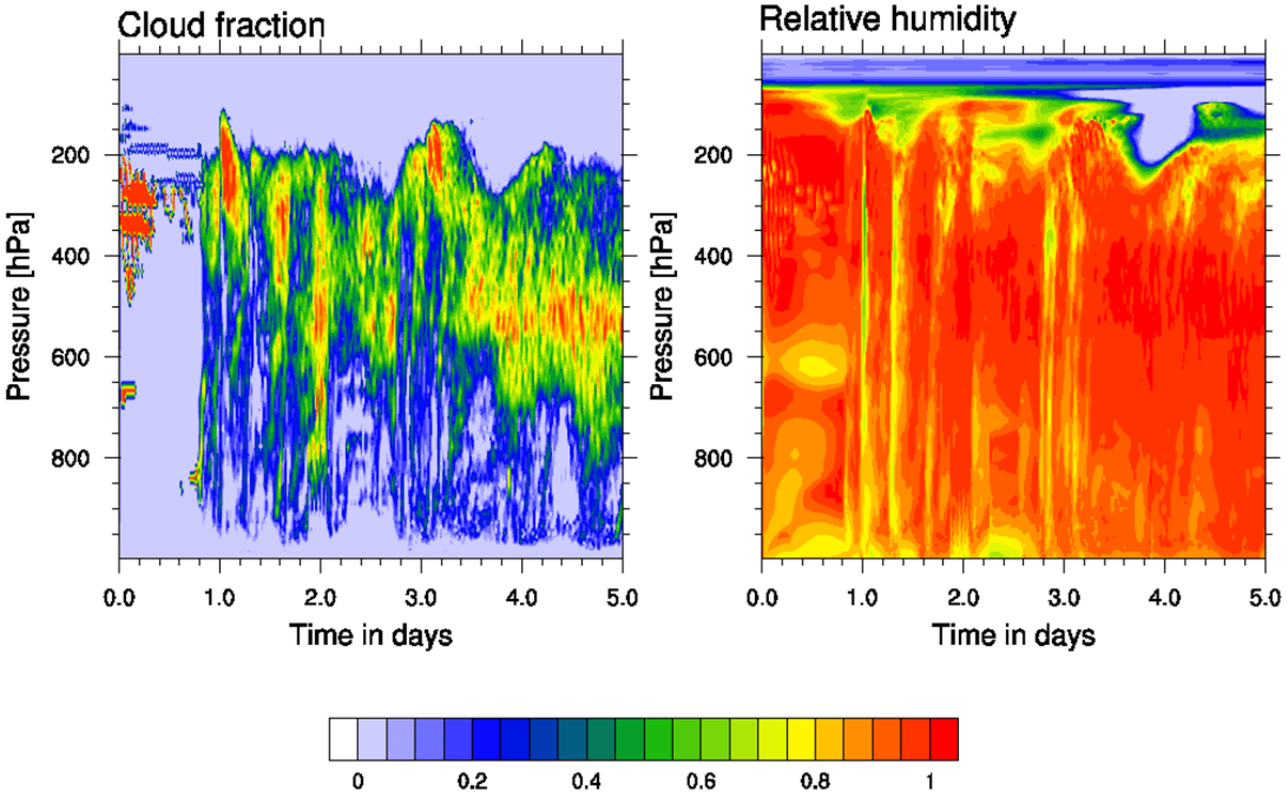
We studied a reduction in precision in the cloud-resolving model (CRM) which is used in the superparametrised version of IFS. The CRM is a limited area model with one horizontal and the vertical dimension that is run within each grid-cell of the global simulation. The CRM is using the prognostic parameters of the global model simulation as boundary conditions. On the other hand, the global model is using information of the CRM to represent sub-grid-scale dynamics down to cloud-resolving scale, in particular the representation of convective processes. To this end, the CRM can be interpreted as an expensive sub-grid-scale parametrisation scheme.
This setup has been tested in the US NCAR climate model before and allowed improvements in the representation of the tropics and in particular the Maddan-Julian-Oscillation and the diurnal cycle in convection globally. However, the use of a CRM in each grid cell increases computational cost significantly and the huge computing cost is the main disadvantage of this setup. The use of reduced numerical precision will potentially be able to reduce the computational cost. The CRM feeds back into the global simulation via averaged quantities only, such that we expect rounding errors of reduced precision hardware to be of minor importance for the global simulation if precision is reduced in the CRM only.
It is non-trivial to find the optimal level of precision that should be used in the CRM. However, if we want to optimise the use of reduced precision hardware beyond single precision, it is not enough to use only one level of numerical precision for the entire CRM. We need to define individual precision levels for small blocks of code or individual parameters to allow the optimal reduction of precision and therefore computational cost. We therefore introduced an automatic search for the optimal precision level of individual parameters. To do this automatic search successfully, we need to define an acceptable level of quality for reduced precision simulations that can be used to decide whether the impact of rounding errors is significant. To this end, we calculate the difference between simulations that reduce an individual precision level, keeping the rest of the model at high precision, against double precision control simulation and compare these differences against the ensemble spread of simulations with slightly perturbed initial conditions. We have studied one approach that identifies the optimal level of precision in the significand of individual parameters and a second approach that used half precision for as many parameters as possible. Results show that we can reduce the amount of bits that is needed to represent the most important fields of the model by 32% for the half precision setup and by 56% for flexible precision setup when compared to simulations in single precision. We have also tested a reduction of precision for the most expensive subroutines. This will generate additional savings that can be estimated to be 9% and 11% for simulations with the superparametrisation setups that is using half precision and the setup with flexible precision for floating point significands respectively.



The Figure above shows results for the cloud fraction within IFS in a grid cell in the tropics for the double precision control simulation, a simulation in double precision that is using a three dimensional CRM (high resolution), and the two model setups in reduced precision. Both setups produce reasonable results. However, we do see some differences close to the surface for the half precision simulation that needs further investigation.

If the optimal level of precision that can be used in simulations is identified using the automated search algorithm outlined above, this information holds interesting implications for studies of model uncertainty. The level of precision that can be used for a specific parameter provides information about parameter uncertainty. We could also show that ensemble simulations can be based on rounding errors that will deliver a reasonable spread even if precision levels higher than single precision are used (it is known that single precision in the CRM is sufficient). Furthermore, we can use the information on numerical precision to reduce the complexity of the model and remove model parts that are hardly used. These model parts can be identified in the precision analysis since a minimal level in precision can be accepted. In the specific model under investigation, the precision analysis revealed that the precision of the turbulent kinetic energy scheme (TKE scheme) can be reduced to only 5 bits in the significand with no strong impact on the dynamics of the CRM. This indicates that the TKE scheme does not have a strong impact on model simulations. Therefore, we tested whether we can remove the TKE scheme from the model setup. A second change to the

model setup will reduce the polynomial order of the saturation curves of water vapour and its derivative from nine to four in an adapted version of the CRM since it was found that the coefficients of the highest polynomial orders can be represented in half precision with no strong impact on model results. However, these coefficients are very small and actually represented by zero in half precision, this is a clear indication that the polynomial order of the saturation curve can be reduced.



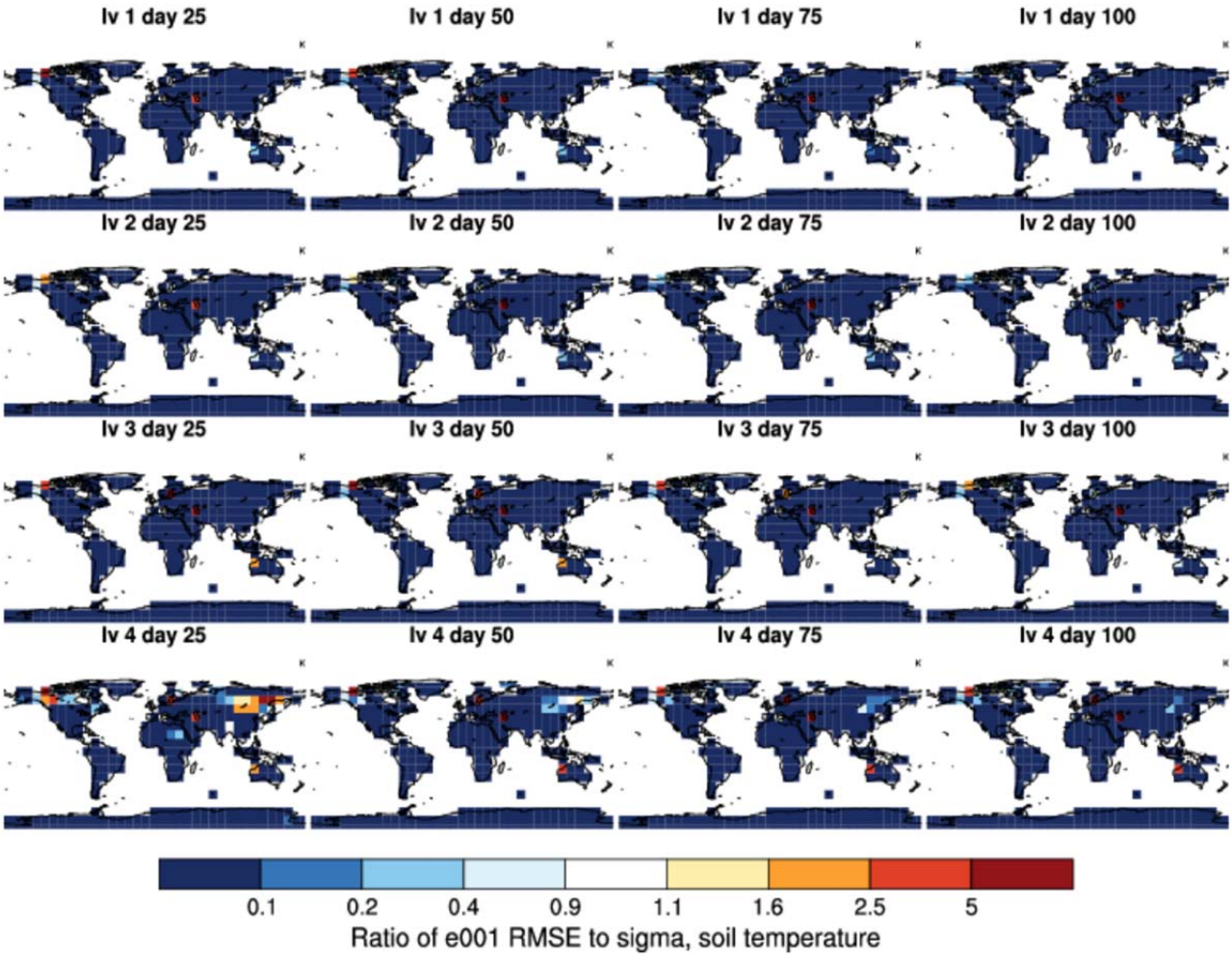
The Figure above shows results of a simulation in double precision with a model setup that does not use the TKE scheme and adjusts the water vapour saturation curve to a lower order. Results are comparable to results from the control simulation (see previous Figure). This confirms that the reduction of model complexity in the co-designed setup did not show a significant change in model behaviour or a strong reduction in model quality. However, the changed version of the superparametrised setup reduced model runtime by ~10%.

The land-surface scheme of IFS in reduced numerical precision

We introduced the emulator for reduced precision to the land surface scheme of IFS. Similar to the cloud-resolving model in super-parametrisation, the land-surface scheme is independent between grid-cells and using the prognostic values of the global simulation as boundary conditions. The prognostic fields of the surface scheme appear to have only limited impact for short-term forecasts of IFS and only field values at the surface interact with the global atmosphere model. Therefore, we expect that numerical precision can be reduced significantly for many quantities in the surface scheme with no strong impact on the global simulation. We compared the impact of a reduction in precision against the spread in ensemble simulations, which acts as an estimate of model uncertainty. If a reduction in precision does not change the spread of the ensemble, we can assume that the use of reduced precision will not degrade results significantly.

The figure below shows the root-mean-square-error of soil temperature if the land surface model is run in 23-bit precision for floating point numbers (measured against a control run at double precision/64-bits) shown as a fraction of the average spread of the ensemble from comparable runs using the coupled model (System 4). Runs are initialised on 1st May and run for four months. Results are shown globally for four snapshots throughout the run, at all four vertical soil levels. It is visible that most of the difference between the reduced precision and the double precision simulation is much smaller compared to the spread of the ensemble simulation, in particular for the

surface layer. This suggests that the model error that is generated by a reduction in precision is not significant for global simulations.

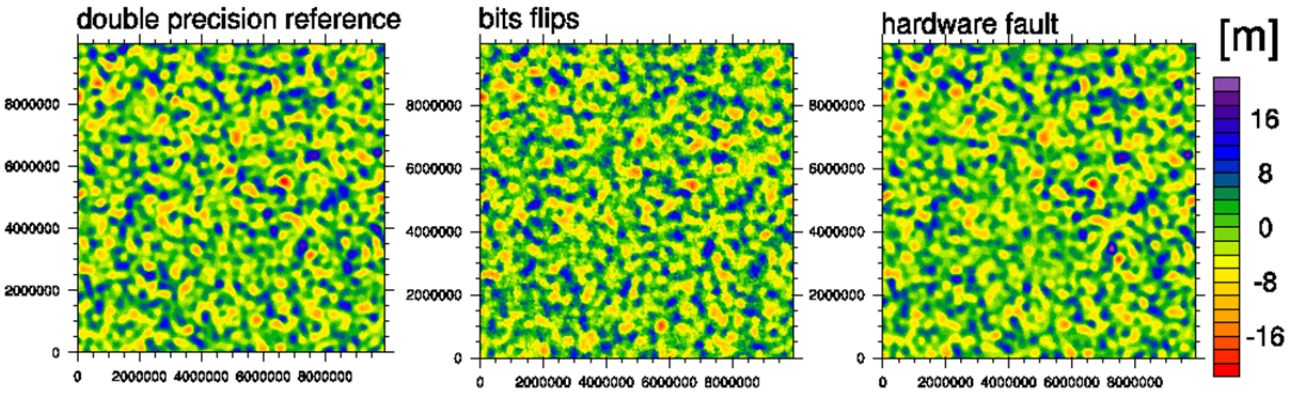


<u>To secure the dynamical core of a weather and/or climate model against hardware faults</u>

One of the most alarming threats for weather and climate predictions on future high performance computing architectures is hardly studied in the community of Earth System modelling yet: The presence of frequent hardware faults that will hit weather and climate simulations as we approach exascale supercomputing.

We worked on an approach to make model simulations resilient against hardware faults using a backup system that stores coarse resolution copies of prognostic variables. Frequent checks of the model fields on the backup grid allow detecting the most severe hardware faults. The prognostic variables on the model grid can than be identified and restored from the backup grid to continue model simulations with no significant delay.

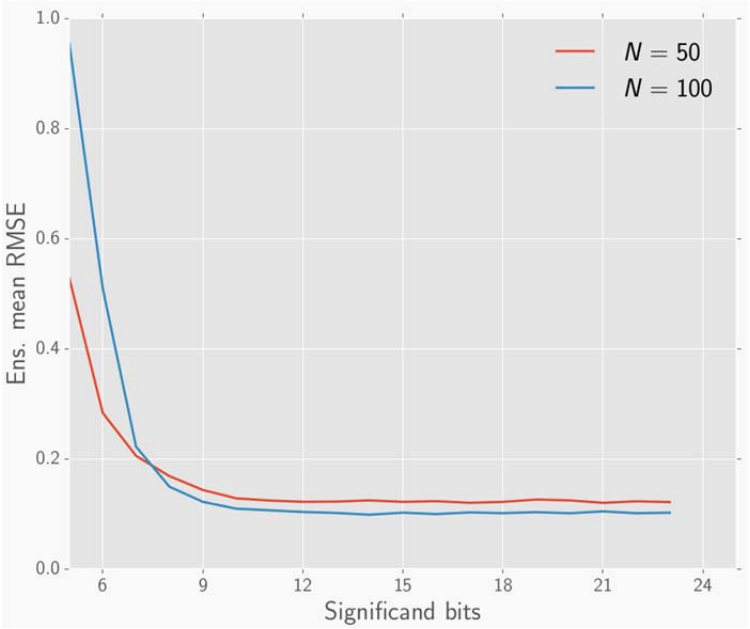
We started to investigate the use of the backup grid in model simulations with a C-grid shallow water model. We emulate frequent bit flips or the loss of information of prognostic parameters in large areas of the domain. As long as the backup system is used, simulations do not crash and a high level of model quality can be maintained. The overhead due to the backup-system is reasonable and runtime is increased by only 13% for the shallow water model.



The Figure above shows the height field for simulations that were initialised with random initial conditions. The last 500,000 timestep of the simulation were performed on hardware with emulated bit flips in floating point operations and floating point assignments (middle) and a simulation that copes with two hardware crashes that set all prognostic variables in 1/16th of the domain to "Not A Number" (NAN) twice within the simulation (right). The bit flips would produce NANs within the simulation after 53 timesteps if no backup system was used. With the backup system in place, both model simulations produce reasonable results with small perturbations compared to the control simulation.

Reduced numerical precision in data assimilation

We started to investigate the use of reduced numerical precision in data-assimilation. In a first approach, we studied an Ensemble Kalman Filter that was implemented in a Lorenz'96 model. Preliminary results suggest that precision can be reduced significantly. The Figure below shows the root-mean-square-error after assimilation plotted against the number of bits that is used in the significand of floating point numbers. Results suggest that precision can be reduced to ~10 bits with no large penalty. This would suggest significant savings that could be reinvested into the use of more ensemble members to improve the assimilation. We will continue our efforts with more complex model setups and aim to investigate reduced precision in data assimilation with Ensemble Kalman Filters within IFS in the future.



A scale-selective approach to precision in IFS

One of the main motivations for the use of inexact hardware in atmospheric modelling is to treat different parts of the atmospheric dynamics with customized numerical precision to reflect their inherent uncertainties. For the atmosphere, it is a useful approach to reduce numerical precision with the considered spatial scales. This is intuitive since small scale dynamics close to the grid scale can hardly be resolved within numerical models. Viscosity often needs to be added to model

simulations to remove kinetic energy, which is building up at the grid-scale due to the turbulent cascade of energy, and tends to smear out small-scale structures in the model fields. Parametrisation schemes that are used to represent sub-grid-scale features generate large uncertainties and have a strong impact on precision at these scales as well. As a result, the quality of the solution at very small spatial scales will not be very good and it can be assumed that they will hardly be affected by rounding errors. On the other hand, contributions of non-linear terms are comparably small for large-scale dynamics at scales of thousands of kilometres and it is much easier to calculate these dynamics at high precision. Therefore, numerical precision should remain large when calculating these scales. Fortunately, most of the computational cost will be caused by the calculation of dynamics close to the grid-spacing (a decrease in horizontal gridspacing by a factor of two will cause an increase in computational cost by approximately a factor of 8 due to a factor of two for each dimension in space and time) and it is most important in terms of forecast quality that large-scale dynamics are calculated correctly.
We continued the study of scale-selective precision in simulations with a model of the surface quasi-geostrophic equations. This model is similar to the IFS in that it is updating the timestep in spectral space and calculating the tendencies of the non-linear terms that are needed for the timestep in grid-point space. First results that reduce precision much stronger for the calculation of the high wavenumbers are promising. We will continue with a similar study of scale-selective precision within simulations with IFS after we have finished the tests in the surface quasi-geostrophic model.

List of publications/reports from the project with complete references

F. Vana, P. D. Düben, S. Lang, T.N. Palmer, M. Leutbecher, D. Salmond, G. Carver (2016), Single precision in weather forecasting models, submitted to Monthly Weather Review

T. Thornes, P. D. Düben, T. N. Palmer (2016), On the Use of Scale-Dependent Precision in Earth System Modelling, submitted to QJRMS

P. D. Düben, 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

P.D. Düben, Parishkrati, S. Yenugula, J. Augustine, K. Palem, J. Schlachter, C. Enz, T.N. Palmer (2015), Opportunities for energy efficient computing: A study of inexact general purpose processors for high-performance and big-data applications, Proceedings -Design, Automation and Test in Europe, 764-769

P. D. Düben, S. Dolaptchiev, 2015: Rounding errors may be beneficial for simulations of atmospheric flow: Results from the forced 1D Burgers equation, Theoretical and Computational Fluid Dynamics, 29, 311–328

P. D. Düben and T. N. Palmer, 2014: Benchmark Tests for Numerical Weather Forecasts on Inexact Hardware, Mon. Wea. Rev., 142, 3809–3829

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

Summary of plans for the continuation of the project

(10 lines max)

We will continue the study of the use of reduced numerical precision to calculate different parts of IFS. For the next steps, we will perform model simulations in larger model setups with both the reduced precision super-parametrisation version of IFS and the reduced precision land-surface scheme. We will also continue the study of reduced numerical precision within an ensemble Kalman filter and approach the use of scale-dependent precision levels with larger model setups with the hope to be able to tackle models of the level of complexity of the full IFS soon. The emulator for reduced numerical precision will remain the main tool to apply reduced precision in the near-term future. Due the large cost for simulations with the superparametrised model setup, we will certainly use up all available billing units until the end of this year.