REQUEST FOR A SPECIAL PROJECT 2017–2019

MEMBER STATE:	United Kingdom

Principal Investigators[1]**:** Tim Palmer and Peter Düben

Affiliation: University of Oxford

Address:

Department of Physics
Atmospheric, Oceanic and Planetary Physics
Clarendon Lab.
Parks Road
Oxford OX1 3PU

E-mail: T.N.Palmer@atm.ox.ac.uk

Other researchers:

Andrew Dawson, Samuel Hatfield, Stephen Jeffress, David MacLeod, Aneesh Subramanian, Tobias Thornes

Project Title:

The use of imprecise arithmetic to increase resolution in atmospheric models

If this is a continuation of an existing project, please state the computer project account assigned previously.	**SP** GBTPIA	
Starting year: (Each project will have a well-defined duration, up to a maximum of 3 years, agreed at the beginning of the project.)	2017	
Would you accept support for 1 year only, if necessary?	YES ☒	NO □

Computer resources required for 2017-2019: (To make changes to an existing project please submit an amended version of the original form.)		**2017**	**2018**	**2019**
High Performance Computing Facility	(SBU)	15.000.000	15.000.000	15.000.000
Accumulated data storage (total archive volume)[2]	(GB)	15.000	15.000	15.000

***An electronic copy of this form must be sent via e-mail to**:* *special_projects@ecmwf.int*

Electronic copy of the form sent on (please specify date): 28th June 2016

Continue overleaf

[1] The Principal Investigator will act as contact person for this Special Project and, in particular, will be asked to register the project, provide an annual progress report of the project's activities, etc.

[2] If e.g. you archive x GB in year one and y GB in year two and don't delete anything you need to request x + y GB for the second project year.

Principal Investigator: Tim Palmer and Peter Düben

Project Title: The use of imprecise arithmetic to increase resolution in atmospheric models

Extended abstract

It is expected that Special Projects requesting large amounts of computing resources (1,000,000 SBU or more) should provide a more detailed abstract/project description (3-5 pages) including a scientific plan, a justification of the computer resources requested and the technical characteristics of the code to be used. The Scientific Advisory Committee and the Technical Advisory Committee review the scientific and technical aspects of each Special Project application. The review process takes into account the resources available, the quality of the scientific and technical proposals, the use of ECMWF software and data infrastructure, and their relevance to ECMWF's objectives. - Descriptions of all accepted projects will be published on the ECMWF website.

Motivation of the proposed research

Global Earth System models are essential to provide meaningful forecasts of future weather and climate, and reliable forecasts of weather and climate are of enormous importance for the preservation and generation of prosperity in society. The quality of a global Earth System model depends on both the used resolution and the complexity of the model. However, resolution and complexity are limited by computational performance.

During the last couple of decades, model developers could trust in an exponential growth of computational power which allowed the use of ever-increasing resolution. However, the increase in performance of individual processing cores is stagnating as we reach physical limits on the size of transistors and energy density for silicon technology. Recently, performance could still be improved using more and more processing cores in parallel. However, it seems that even an excessive use of parallelism will not keep up the exponential growth in computing power in the near term future. Furthermore, a strong increase in the numbers of processors that are used in parallel for model simulations will cause a strong increase in power consumption, which is already a significant cost factor for today's high-performance-computing centres. Today, Earth System models are already running on some of the fastest supercomputers of the world but it is still not possible to resolve important processes, such as convection in the atmosphere and mesoscale eddies in the ocean, in operational medium-range weather forecasts or global climate simulations.

This project will continue our study of reduced numerical precision in simulations of weather and climate models. If numerical precision can be reduced in simulations, power consumption of supercomputers can be reduced and performance can be increased. Furthermore, a reduction in precision will also allow a strong reduction of the amount of data that needs to be stored. The cost of data storage is larger for weather and climate models in comparison to most high-performance computing applications. If cost savings are reinvested into higher resolution or an increased number of ensemble members, weather and climate predictions can be improved.

There are several ways to reduce numerical precision and to trade precision against performance in simulations of weather and climate models. The simplest way to reduce precision in state-of-the-art models is to reduce precision from double precision (64 bits) to single precision (32 bits). This is possible on existing high-performance computing hardware and will allow a performance increase by up to a factor of two. Precision could also be reduced to half-precision (16 bits) for example on Pascal GPUs from NVIDIA. A reduction of precision in data storage can also be realised on existing high performance computing hardware. Another approach to trade numerical precision against performance is to allow occasional hardware faults within model simulations. Frequent hardware failures are already reality for existing supercomputers. For example, the Los Alamos National Laboratory's ASCI Q machine experienced 27.2 CPU failures per week (see Michalak et al. 2005). If, on the other hand, a small rate of hardware faults can be accepted in large HPC applications, computing cost can be reduced significantly since the requirement to produce the correct result under all circumstances is taking up a huge amount of resources. The trade-off to reduce computing cost and allow a certain amount of faults was, for example, discussed for stochastic processors that overscale the voltage which is applied to the processing core. This will cause occasional bit flips but allow significant savings (see for example Kahng et al. 2010).

Results of the study of reduced numerical precision have the potential to lead to a paradigm change for numerical precision in atmospheric modelling around the world. Since reduced precision hardware has never been tested in an application as big as an entire global weather forecast model, results of this study will be relevant for many other high performance computing applications, especially in the field of computational fluid dynamics. Results will also be essential for future developments of reduced precision hardware.

Summary of the most relevant work in previous studies in our group

We have studied the use of single precision in the OpenIFS model. IFS used double precision as a fixed precision level in the entire model for the last decades. However, we could show that it is possible to switch back to single precision in almost the entire model and that simulations in double and single precision produce results of similar quality (see Düben and Palmer, 2014). 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, following the same changes that we have performed for OpenIFS, and introduced a switch to choose between single and double precision in the latest IFS cycle. Results with the IFS model confirm the result of our evaluation with 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 (Vana et al. submitted to MWR 2016). The use of single precision appears to be a promising candidate to reduce computational cost of operational forecasts significantly. However, further tests at higher spatial resolution will still be necessary to verify whether simulations at single precision at operational forecast resolution are degraded in comparison to double precision simulations. If these tests are successful and if single precision is used in the operational forecast in the future, savings in computing cost will easily exceed the cost of this special project proposal.

We have also studied the use of emulated reduced precision in different parts of the OpenIFS model. 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).

This emulator allowed the study of reduced numerical precision within the land-surface model and the cloud-resolving model that is used in the superparametrised model setup of IFS. Results show that numerical precision can indeed be reduced significantly beyond single precision. For the cloud resolving model of the superparametrised setup, we can reduce the amount of bits that is needed to represent the most important fields of the model by 32% if we use half precision for as many real numbers as possible and by 56% if we use flexible precision in the significand of floating point numbers with no strong reduction in model quality. For the land-surface model, we can reduce precision to 23-bit precision for floating point numbers and see a difference between the reduced precision and the double precision simulation that is much smaller compared to the ensemble spread for almost the entire globe and in particular in the important surface layer. For both projects, we are in the process of writing scientific papers that will be submitted very soon. We have also performed numerous tests with reduced numerical precision in toy models, such as Lorenz'96 or a Burgers equation model, and a spectral dynamical core (IGCM; see for example Düben et al. JCP 2014 and Düben and Palmer MWR 2014).

Projects that will be investigated in the next three years

To continue the evaluation of the superparametrised model setup and the land surface scheme:

The work on the use of reduced numerical precision in the superparametrised model setup and the land surface scheme of IFS will be continued. We aim to perform global simulations with the OpenIFS model at reasonable resolutions (such as T159, 91 vertical levels) with both the co-designed superparametrised setup, that is using a reduced model complexity for the cloud resolving model, and simulations with emulated reduced numerical precision in the cloud resolving model. The co-designed model setup used information from a detailed precision analysis to identify parts of the model that do not have a strong impact on model dynamics to reduce model complexity.

Simulating extreme weather events at reduced numerical precision:

We will test how a reduction in numerical precision will influence the predictions of specific extreme weather events with a specific focus on tropical cyclones. First tests will compare model simulations in single precision with model simulations at double precision and test whether results with single precision are degraded. In a second step, we will continue the investigation using the emulator to reduce numerical precision in large parts of the model setup. The emulator may also be used to mimic the use of specific configurations of inexact hardware within model simulations.

Scale dependent precision in weather and climate predictions:

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 that reflects their inherent uncertainties. For the atmosphere, it is a useful approach to reduce numerical precision with 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 that 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. In contrast, 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 grid-spacing 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 have already studied this approach in toy models and a spectral dynamical core (Thornes et al. submitted to QJRMS 2016, Düben et al. JCP 2015, Düben and Palmer MWR 2016). Results show that the use of scale-dependent precision is promising a very strong decrease in precision for the most expensive parts of the model simulations. We will continue this study within the OpenIFS model. In particular, we will investigate the use of scale dependent precision in calculations of the time-stepping scheme in spectral space as well as the Fast Fourier and the Legendre Transformation.

Data assimilation in reduced numerical precision:

We already 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 to ~10 bits in the significand of floating point numbers with no large penalty. This would allow 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 assimilation with Ensemble Kalman Filters within IFS in the second and third year of the proposed project.

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

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.

We will continue to investigate methods to allow model simulations in the presence of hardware faults. This will also include tests that investigate how model simulations with OpenIFS will be influenced by bit-flips. Soft errors can be simulated using the emulator for reduced numerical precision such that this study can be based on the reduced precision OpenIFS model setups of the other projects.

Justification of computer resources and technical characteristics

We will perform model simulations that emulate the use of reduced numerical precision in large parts of the OpenIFS model and we will also continue the analysis of model simulations with OpenIFS that use single precision. To compare the error in model simulations due to rounding errors with model uncertainty and model error, we will need to evaluate results for ensemble simulations. A thorough test of a model setup that is using reduced numerical precision would need ensemble simulations at T159 horizontal resolution with 91 vertical levels. The ensemble simulations for a forecast of ten days shall be calculated with fifty ensemble members for thirty different forecasting dates, to get sufficient statistics. To calculate such a setup in double precision with stochastic parametrisation would need approximately 60,000 System Billing Units (SBUs). While the use of the single precision setup would reduce computing cost by up to 40%, the use of the emulator for reduced precision floating point arithmetic will, unfortunately, reduce model performance significantly by approximately a factor of 20, at least for those parts of the model that are run with emulated reduced precision. If we perform the same set of ensemble simulations at T159 resolution, we will therefore need approximately 1,200,000 SBUs. We will need to test several different setups in which the emulator is used in different parts of the model for the different projects that are outlined above and at different precision levels or fault rates. Some of the project, in particular the simulation of extreme weather events, would not need ensemble simulations at different start dates but rather simulations at higher spatial resolution. Simulations with the superparametrised model setup will be much more expensive compared to the standard model setup. To this end, we will not be able to perform those simulations for many start dates as well. To provide a realistic approximation for the SBUs that are needed within one year, we multiply the cost for a single ensemble simulation that is using the emulator for reduced precision with an additional factor of 12.5 and end up with the approximated amount of 15,000,000 SBUs per year. Since we have already experience with OpenIFS and the ECMWF supercomputing facilities, we can already start large simulations on the supercomputing facilities within the first year of the project and will not need more SBUs in the second and third year.

References

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

A. Kahng, S. Kang, R. Kumar, J. Sartori, 2010: Slack redistribution for graceful degradation under voltage overscaling, Design Automation Conference (ASP-DAC), 15th Asia and South Pacific, 2010, pp. 825 –831.

S. E. Michalak, K. W. Harris, N. W. Hengartner, B. E. Takala, and S. A. Wender, 2005, Predicting the number of fatal soft errors in Los Alamos National Labratory's ASC Q computer. IEEE Transactions on Device and Materials Reliability, 5:329–335.

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

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