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	Jan 2014 – July 2014
Project Title:	The use of imprecise arithmetic to increase resolution in atmospheric models
Computer Project Account:	spgbtpia
Principal Investigators:	Prof. Tim Palmer, Dr Peter Düben
Affiliation:	University of Oxford
Name of ECMWF scientists collaborating to the project	Dr Glenn Carver, 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)	0	0	5000000	651836
Data storage capacity	(Gbytes)	0	0	6000 GB	~100 GB (ECFS only)

Summary of project objectives

(10 lines max)

The aim of this project is to study the limits and prospects of the use of imprecise hardware in a global atmosphere model (IFS). In particular, we want to investigate the potential use of inexact hardware and reduced precision floating point arithmetic in a model of the global atmosphere. 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. A reduction of computing cost allows higher computing power within the same budget for computing facilities. Higher computing power allows higher resolution and the resolution in state-of-the-art atmospheric simulations is still far from being adequate. If the results of this project show that it is possible to use inexact hardware to perform large parts of the computation of an atmosphere model, this project might lead to a fundamental change in future computing hardware used for weather and climate modelling.

Summary of problems encountered (if any)

(20 lines max)

None.

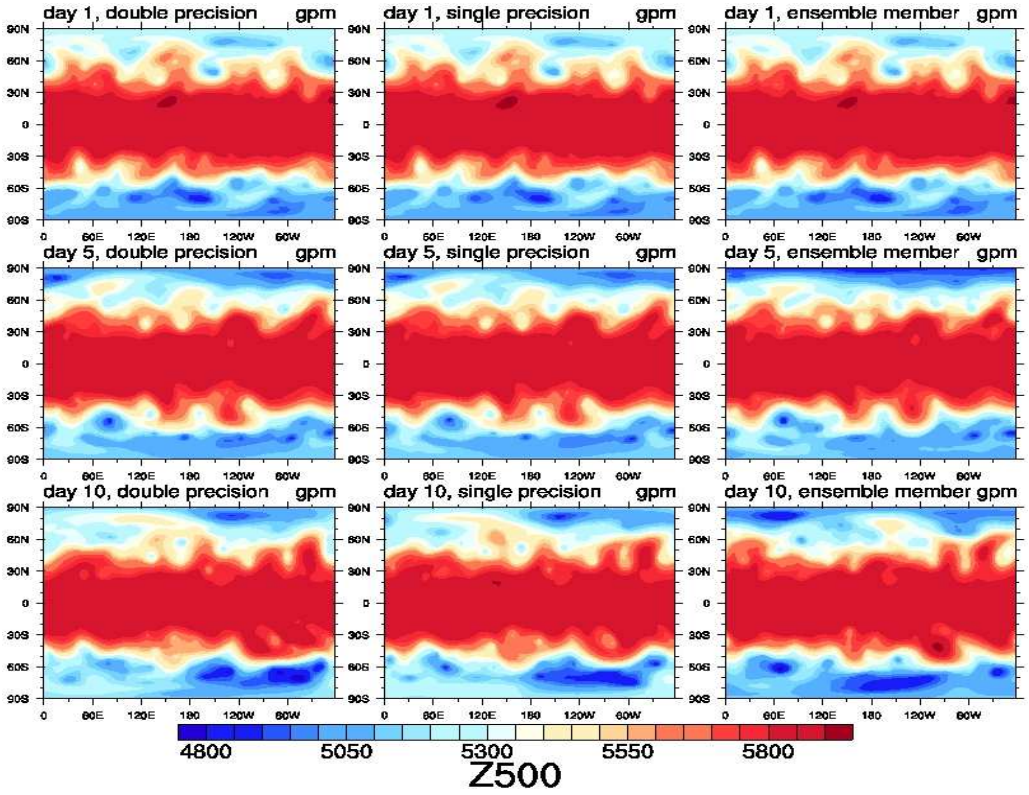
Summary of results of the current year (from July of previous year to June of current year)

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

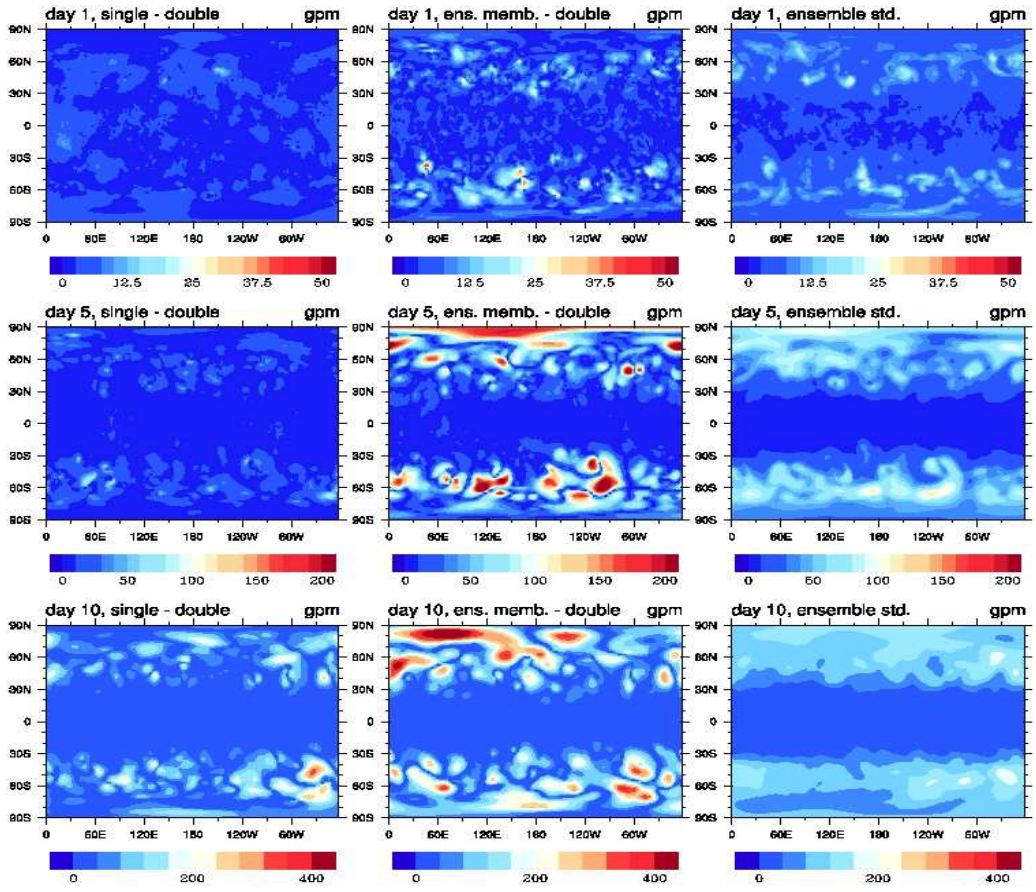
The project is still in a very early stage since it only started half a year ago. Therefore, the computational budget was not yet used in the right proportion. However, we will certainly need more computational power during the second half of the first year for long term, high resolution simulations with OpenIFS in single precision. We will also use the provided computational resources to perform simulations with emulated inexact hardware inside the OpenIFS.

Within the first half year of the project, we focused on the development of a single precision model version of the OpenIFS. Simulations with single precision would reduce the amount of data that needs to be transported and processed by a factor of two, in comparison with double precision arithmetic, and would also speed-up the processing of vectorised loops by a factor of two. OpenIFS has about 500,000 lines of code in more than 2,500 files and a switch from double to single precision represents a serious effort, however a re-writing of large parts of the model code is not necessary. We were already able to develop a running version of OpenIFS that is using single precision for almost all parts of the model. However, we tested only a small subset of possible diagnostics yet and we do not claim that the developed single precision version is working perfectly for all situations at the current stage. It is likely that further changes will be necessary to guarantee stability in long term simulations.

To test the single precision setup, we performed several simulations at different horizontal resolutions (T21, T159, T255, T511, and T639). We present results for a ten day forecast with the highest resolution, T639, which is the resolution of the operational ensemble prediction system at ECMWF within this report. The initial date of the forecast is 1st November 2012, 00:00 am coordinated universal time.



The Figure above shows the results for geopotential at 500 hPa after one, five, and ten days of simulations for the single precision setup, the double precision control simulation of the ensemble forecast performed with the Integrated Forecasting System (IFS) and one ensemble member.



The second Figure shows the difference between either the single precision simulation or one ensemble member of the ensemble forecast compared to the double precision control simulation for geopotential at 500 hPa and the standard deviation of the ensemble forecast. Differences between

the single precision and the control simulation are always smaller than differences between the control simulation and one ensemble member. Differences for the single precision run are reasonably small compared to the ensemble standard deviation. This is promising since the ensemble spread is setup to resemble the uncertainty of the forecast. As long as the single precision simulation stays within the "envelope" of the ensemble system, we can assume that the accuracy of the single precision forecast is approximately equal to the accuracy of the double precision forecast.

All simulations were performed on 64 bit CPU architecture. If a single precision simulation is run on 64 bit architecture and if no vectorisation is used, the flop rate is not necessarily increased compared to the flop rate at double precision. A speed-up with single precision can only result from the reduced amount of data that needs to be stored, transported, or fitted into memory and cache. However, if the speed of a simulation is limited by data transport or memory, we expect an increase in performance by a factor of up to two. For OpenIFS, the measured speed-up was heavily dependent on the computing architecture used and the amount of parallelisation. We obtain hardly any speed-up for a T159 or a T255 simulation on a desktop computer with four MPI tasks (Intel Core i7-3770 CPU @ 3.40GHz x 8 with 15.6 GiB memory), while we see a reduction of computing time by 22%, 27% and 25% when running a T159, T255 and T511 model on one computing node of a CPU cluster (Intel Xeon CPU E5630 @ 2.53GHz with 50 GiB memory) using eight MPI tasks. We could run a T511 single precision simulation on the desktop computer and a T639 single precision simulation on the computing node while this was not possible in double precision, due to the limited memory. Work is in progress to obtain meaningful comparisons within an operational environment of the ECMWF supercomputers. We will also perform tests with different compilers (we used the GNU compiler for the tests above) with more aggressive vectorisation, since we expect that this will allow a larger speed-up.

We also started to incorporate an emulator for reduced inexact hardware into parts of the OpenIFS. We plan to use the emulator to mimic the use of pruned hardware setups which will be developed in cooperation with the working group of Prof. Krishna Palem from Rice University. In "pruning" the physical size of a floating point unit will be reduced by removing parts that are either hardly used or do not influence significant bits in the results of floating point operations. Pruned hardware is promising a strong reduction in power consumption and a significant performance increase. The developed hardware setups will not be available as real hardware due to the huge cost for manufacturing prototype chips. However, it is possible to measure the exact error pattern as well as the savings in power consumption and computing time. The error pattern will be fed into the emulator that is used to mimic reduced precision hardware inside OpenIFS. We recently published a paper on the use of pruned hardware for the Lorenz '95 model (Düben et al. (2014)). The results are very promising and we will now proceed to the larger setup of OpenIFS. Here, we will try to continue the approach of scale separation for which numerical precision is reduced with decreasing spacial scale (increasing wave number).

References

Peter D. Düben, Jaume Joven, Avinash Lingamneni, Hugh McNamara, Giovanni De Micheli, Krishna V. Palem and T. N. Palmer (2014), On the use of inexact, pruned hardware in atmospheric modelling*, Phil. Trans. R. Soc. A*

List of publications/reports from the project with complete references

Düben, Peter D. and Palmer, T. N., Benchmark tests for numerical forecasts in inexact hardware, under review in Monthly Weather Review

Düben, Peter D., Can rounding errors be beneficial for weather and climate models?, in preparation.

Summary of plans for the continuation of the project

(10 lines max)

We want to continue the investigation of the single precision version of OpenIFS. In particular, we want to simulate higher resolution setups (e.g. T1279) for longer integration times (e.g. one year) and different starting dates. A more detailed analysis of important diagnostics will be necessary since it is likely that local code will still need to be changed to avoid erroneous long term behaviour and divisions by zero values. We will also start to emulate the use of inexact hardware within the OpenIFS model. We will focus on cost intensive parts of the model, such as the cloud scheme, the FFT, and the Legendre Transformation. Here, the main challenge will be to apply the approach of scale separation, which has shown to be very successful in spectral dynamcial cores, within the complex OpenIFS model code. We will probably work together with hardware developers, namely the group of Prof. Krishna Palem from Rice University, to develop customised processors for the use in OpenIFS.