SPECIAL PROJECT FINAL REPORT

All the following mandatory information needs to be provided.

Project Title:	The use of imprecise arithmetic to increase resolution in atmospheric models
Computer Project Account:	spgbtpia
Start Year - End Year :	2017 - 2019
Principal Investigator(s)	Tim Palmer
Affiliation/Address:	Atmospheric, Oceanic & Planetary Physics University of Oxford Clarendon Laboratory, University of Oxford, Sherrington Rd, Oxford OX1 3PU
Other Researchers (Name/Affiliation):	Matthew Chantry – University of Oxford Peter Dueben – ECMWF

The following should cover the entire project duration.

Summary of project objectives

(10 lines max)

This project aimed to assess the possible value of reduced numerical precision to produce more accurate weather and climate forecasts. This would be made possible by exchanging reduced numerical precision, and the computational savings that it brings, for increased model complexity, e.g. increased model resolution. In a previous special project we had shown the value of moving from double precision to single precision, which is in the process of being adopted at ECMWF as a result of our research and continuing work at ECMWF. Here we investigate the possible value of even further reductions, in particular half-precision. This has huge theoretical benefits as computing in general is moving towards heterogenous hardware (e.g. a mix of CPU and GPU computing) and GPU units are now commonly equipped with half-precision number formats. This special project allows our investigation to assess the benefit to the IFS, including simulations at up-to operational forecast resolutions.

Summary of problems encountered

(If you encountered any problems of a more technical nature, please describe them here.)

No major problems encountered. We would like to thank Paul Dando for his help with many small issues and general help ensuring the smooth running of this project.

Experience with the Special Project framework

(Please let us know about your experience with administrative aspects like the application procedure, progress reporting etc.)

We have found the special project framework to be a transparent and fair process.

Summary of results

(This section should comprise up to 10 pages, reflecting the complexity and duration of the project, and can be replaced by a short summary plus an existing scientific report on the project.)

Half-precision is a floating-point number format that uses 16bits to represent each number, compared with 64bits used for double-precision numbers. This corresponds to a representation error (or rounding error) of 10^{-3}, compared with 10^{-16} for double precision or 10^{-6} for single precision. This larger representation error, alongside the smaller dynamic range (the smallest to largest representable numbers) means that attempting to introduce half-precision throughout the IFS codebase (or any other significant weather or climate code) would lead to very poor results and no obvious route to find or understand the problems introduced. Instead we approach the task by breaking a weather code into kernels, self-contained sets of subroutines and functions, which are significant in cost but small enough to understand the results. This approach has also been used at ECMWF with the "dwarves" in their ESCAPE scalability project (Muller 2019). The possible benefits of moving a significant proportion of the calculations to half-precision are notable. A repeat of the 40% increase in computational efficiency seen in the move from double- to single-precision (Vana el al. 2017) could be expected. Or, in areas using dense matrix-matrix multiplications, factors of up to 16 times faster than double-precision are possible (NVIDIA 2017). This would require ECMWF (or other centres) to invest in heterogeneous systems featuring GPUs, but this appears to be the generic direction of large-scale computing. In this report we shall summarise the results of two papers published by our group which benefit from this special project. Both of these investigate the possibility of using half-precision in kernels of the IFS code. Beyond these two papers, we have a body of work assessing other kernels of weather and climate forecasting, but these have not yet been scaled up to systems needing the computational power given by special project resources.

First, we will summarise the results of Chantry et al. (2019) which considered the impact of reduced numerical precision in the spectral space calculations. The spectral space calculations are an important kernel of the dynamical core, being the location of the implicit pressure solve in this formuation. While the costs of this kernel are not dominant (typically costing ~5% of the run-time cost) this kernel poses interesting questions about the importance of precision (and thus error and uncertainty) as a function of lengthscale. In the spectral space calculations, the prognostic fields are represented as coefficients of Legendre and Fourier modes, which occurs to facility the linear solve for the pressure field. These coefficients and their related modes represent different horizontal lengthscales, ranging from mean field (synoptic scales) down to grid-scale length fluctuations. These scales have different predictability timescales and varying levels of constraint from the data-assimilation process. Our thesis was that the largest scales should therefore require higher precision and smallest scales require the lowest. This is born out in figure 1, which shows the number of bits required to represent the significand (i.e. the precision) as a function of the spectral wavenumber. Largest scales require up to 15 significant bits to satisfy a threshold in L2-norm error in our T255 simulations with OpenIFS. The smallest scales required as few as 4 bits (corresponding to a relative error of ~0.0625) to produce an accurate forecast. For all total wavenumbers greater than 10, fewer than 10 significant bits are required for an accurate forecast. This is notable as this would mean that half-precision could be considered for these modes. It should be noted that the issue of the dynamic range of half-precision was not considered in this study. However, as the spectral space calculations are linear, a rescaling could be used to keep these calculations with the dynamic range used in half-precision. For the higher horizontal resolution simulations of T511 we found that as many as 99.98% of coefficients (and hence calculations) could be carried out at half-precision. These results also provide motivation to assess the importance of precision to the spectral transforms. The spectral transform kernel sits side either side of the spectral-space calculations and have a far more significant cost.

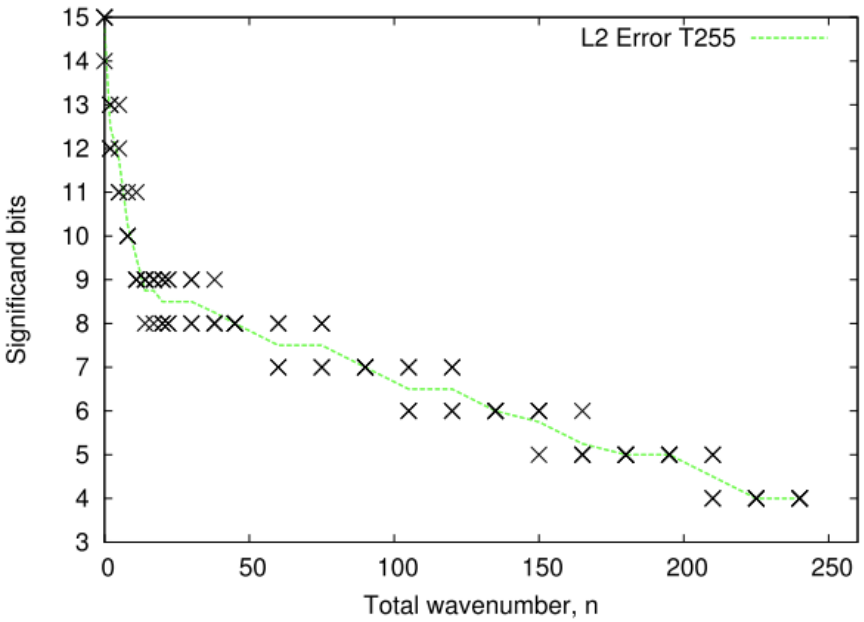

Figure 1: Required significant bits in the spectral space calculations, as a function of total wavenumber, to meet an L2-norm threshold for T255 resolution simulations. Large wavelength and scales (low n) require more significand bits while smaller scales (high n) require fewer bits. See Chantry et al. (2019) for further details.

Next we summarise the results of the Hatfield et al. (2019) which used this special project for its computing. This paper won the best paper prize at PASC2019 (Platform for Advanced Scientific Computing) conference. Here, we built upon the work in Chantry et al. (2019) which established a clear scale-dependence of the precision required for the spectral coefficients. We take the Legendre transforms kernel, which alongside the Fourier transforms to form the spectral space transformations. The Legendre transforms have the worst scaling as resolution is increased, with the cost of this component growing as n^3, where n is the number of points in the horizontal directions. Excluding the cost of IO (input/output operations) the transforms currently account for 41% of the total IFS cost. In recent experiments on Summit running the non-hydrostatic IFS at 1.45km (again without IO) this is increased to 55% of the cost (P.Dueben private communication). Accelerating this kernel would therefore have a major impact on overall cost of running the IFS. The Legendre transforms are essentially matrix-matrix multiplications, a key element of machine learning and a focus area for recent developments in GPU related technology. One example is the V100 NVIDIA GPU which features a matrix-matrix multiplier known as “Tensor Core”. This boasts a speed-up of 16 times when using half-precision over double-precision for such calculations. In addition, the half-precision “Tensor Core” multiplication uses single precision to accumulate the values involved in each dot-product, offering higher precision than a purely half-precision version. This can be crucial in large matrices where increments could otherwise be rounded to zero during accumulation. In our paper we investigated how useful this would be to the IFS code, using the knowledge learnt from our scale-separation paper. In our work we use an emulator to carry out the half-precision (and tensor core) calculations (Dawson & Dueben 2017). This results in simulations that are slower but allows us to test kernels of the IFS without translating them into CUDA.

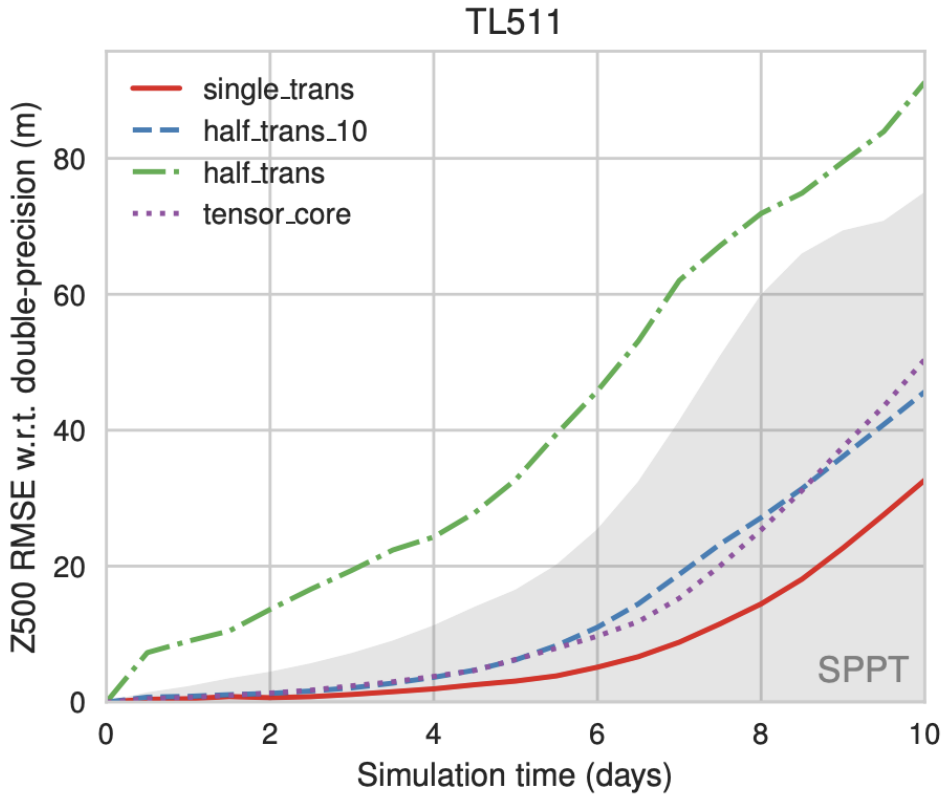

Figure 2: Root-mean-square error for 500 hPa geopotential height field for various experimental setups with respect to a double-precision control experiment at TL511 resolution. Single_trans uses single-precision for the Legendre transform (LT), half_trans uses half-precision for all transforms except the m=0 mode, which is kept at double-preicision, half_trans_10 uses half precision for all modes m>9, tensor_core uses "tensor core" simulated LT for all modes except m=0. (See Hatfield et al. 2019 for more details).

In figure 2 we show results from simulations at T511, emulating the effects of using half-precision for the Legendre transforms. We find that only the m=0 transform, i.e. 1 out of 512 transforms, needs to be carried out at double-precision to still produce a forecast within the uncertainty guideline of SPPT. This is enabled by the "tensor core" accumulation being carried out at single precision. Emulating fully half-precision transforms we find that slightly more of the calculations require double-precision to meet our forecast threshold. To measure the applicability of our results we carry out further tests at T1279, the current operational resolution of the IFS. In figure 3 we show tracks of hurricane Irma as reforecasted in double-precision and our emulated "tensor core" results, where again only the m=0 transform is carried out at double-precision. There is little notable effect on the hurricane track, particularly in the context of the analysis. In Hatfield et al. (2019) further comparisons of forecast quality can be found which consistently show good predictive performance using the emulated tensor core transforms.

The next step in this project will be to realise these benefits by re-coding the current Legendre transform using CUDA to enable the use of GPUs. This promises to be particularly important as ECMWF examines the benefits of moving to a non-hydrostatic dynamical core, which requires an increased number of transforms each timestep. Half-precision Fourier transform packages are also now available and should be studied in the context of IFS to maximise performance improvements.

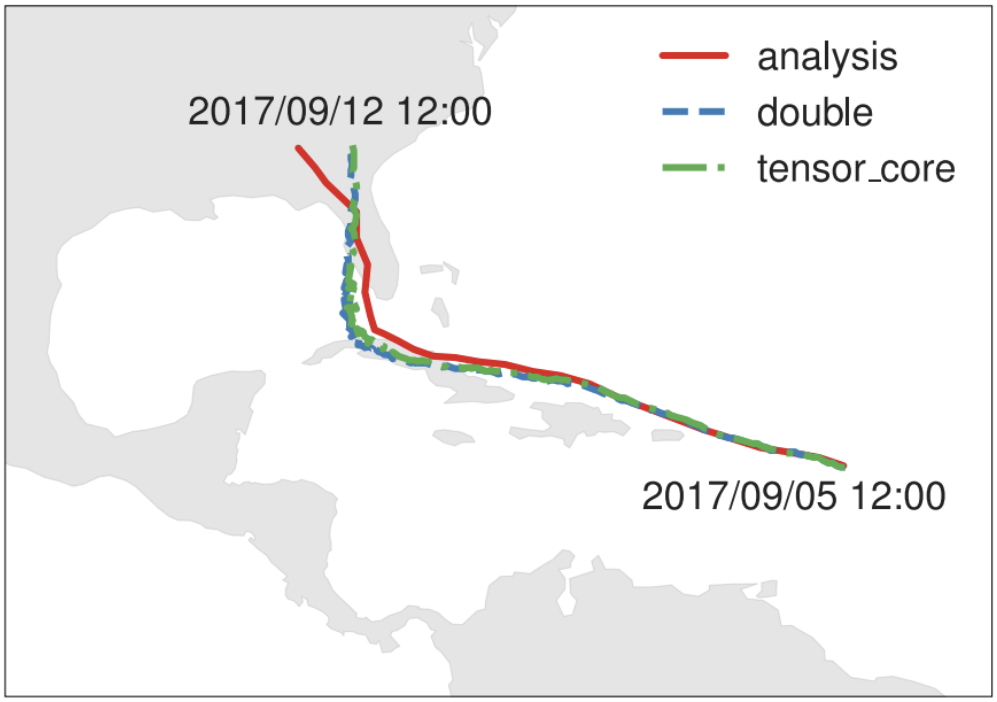

Figure 3: Forecasts of track of hurricane Irma up to day 7 for double-precision and emulated "tensor core" used for the Legendre transforms. Both perform similarly well when compared with the analysis.

To summarise, we have presented two papers which investigate the possible value of half-precision to improving the efficiency of the ECMWF IFS model. We find that in both kernels investigated half-precision cannot be used for all calculations, but guided by science it can be used for many of the calculations. We hope to continue this fruitful line of research in our follow-up project which looks not only at using half-precision but also machine learning as approximate computing for kernels of the weather forecasting system.

Bibliography

Müller, Andreas, et al. "The ESCAPE project: Energy-efficient Scalable Algorithms for Weather Prediction at Exascale." *Geoscientific Model Development Discussions* (2019): 1-50.

Váňa, Filip, et al. "Single precision in weather forecasting models: An evaluation with the IFS." *Monthly Weather Review* 145.2 (2017): 495-502.

NVIDIA. 2017. NVIDIA Tesla V100 GPU Architecture. Technical Report. http://www.nvidia.com/content/gated-pdfs/Volta-Architecture-Whitepaper-v1.1.pdf

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

Dawson, Andrew, and Peter D. Düben. "rpe v5: an emulator for reduced floating-point precision in large numerical simulations." *Geoscientific Model Development* 10.6 (2017).

Hatfield, S., Chantry, M., Düben, P. and Palmer, T., 2019, June. Accelerating high-resolution weather models with deep-learning hardware. In *Proceedings of the Platform for Advanced Scientific Computing Conference* (pp. 1-11).

List of publications/reports from the project with complete references

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

Hatfield, S., Chantry, M., Düben, P. and Palmer, T., 2019, June. Accelerating high-resolution weather models with deep-learning hardware. In *Proceedings of the Platform for Advanced Scientific Computing Conference* (pp. 1-11).

Future plans

(Please let us know of any imminent plans regarding a continuation of this research activity, in particular if they are linked to another/new Special Project.)

This special project has been renewed for the period 2020-2022, spgbtpia. We shall be continuing our research into reduced numerical precision to accelerate weather forecasting. This will include the use of machine learning which can be viewed as an accelerator for uncertain kernels with a weather forecasting system. In particular we will investigate reduced numerical precision and AI emulation for the physical parameterisation schemes, which are inherently uncertain. This work has already begun under the above project.