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	2017
Project Title:	The use of imprecise arithmetic to increase resolution in atmospheric models
Computer Project Account:	spgbtpia
Principal Investigator(s):	Tim Palmer
Affiliation:	University of Oxford
Name of ECMWF scientist(s) collaborating to the project (if applicable)	Peter Duben Antje Weisheimer
Start date of the project:	2017
Expected end date:	2019

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)			15000000	4537795
Data storage capacity	(Gbytes)				Low

Summary of project objectives

(10 lines max)

Investigate the possible benefits of reduced and variable numerical precision on weather and climate prediction. Double precision is used as the default precision level in most weather and climate codes, yet these models contain large sources of uncertainty and error. We are investigating if this high precision is necessary for an accurate forecast. Tests with an OpenIFS version that uses reduced precision in spectral space show good predictive power for hurricane Sandy even with large precision reduction. By using an approach that varies with lengthscale we find that precision can be further reduced.
Single precision superparameterization has also been investigated on this grant. Compared with stochastic parameterised runs we find an increase in skill for forecasts of tropical convection during MJO events but deterioration in some large-scale dynamic fields.

Summary of problems encountered (if any)

(20 lines max)

No major problems have been encountered to date.

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

Double precision, 64 bits to represent each number, is the industry norm for weather codes, as in many other fluid dynamic applications. However, weather and climate codes are unable to resolve all the necessary lengthscales, timescales and physical processes involved in the true system. Instead these are parameterised, often with stochastic elements. When working with an imperfect model it is unclear whether double precision is required. In range of simplified weather models it has been shown that double precision is an unnecessary computational burden [Düben et al., 2014, Düben and Palmer, 2014, Thornes et al., 2017]. This motivated the testing of the Integrated Forecast System (IFS) weather model in single precision [Váňa et al., 2017]. This reduction gave an average computational gain of 40% over double precision, while showing no noticeable degradation in forecast skill. Under this special project, our main goal is to investigate how much further of a reduction in precision is possible and whether a scale-selective approach is beneficial. We will make estimations as to the possible cost savings from further precision reductions.

The uncertainty involved in weather models has a clear lengthscale dependance. At small scales, subgrid-scale processes, parameterisation schemes and data assimilation can introduce large errors. Whereas at large lengthscales, all of these features have a heavily reduced impact and errors are expected to be small relative to the Navier–Stokes dominated dynamics. This variation in uncertainty suggests that the optimal approach would have numerical precision dependant upon the lengthscales involved in each operation. It is this concept that we wish to test, in a full complexity model.

To investigate this scale-selective idea we will consider the OpenIFS model. This paired-down version of the IFS model will facilitate an easier introduction of our reduced precision. Generally, the advantage of a spectral dynamical core is the ease of enacting implicit time-stepping schemes for the linear terms. The resulting elliptical PDEs can be easily solved as the horizontal discretisation is largely separable. The

disadvantage of spectral schemes is the global communication necessary for the spectral transforms. This communication is seen as the harbinger of doom for spectral schemes in moving to exascale computing. However recent algorithmic improvements in the Legendre transformations give hope for their longer-term survival [Wedi et al. 2013]. For our investigation a spectral dynamical core is key to investigating scale-selective numerical precision. On the sphere the typical horizontal spectral decomposition of a field X is

$$X(\theta,\lambda) = \sum_{m=-N}^{N} e^{im\lambda} \sum_{n=|m|}^{N} X_n^m P_n^m(cos(\theta)),$$

where θ and λ are the latitude and longitude. N is the truncation level and n is the total wavenumber. The total wavenumber controls the lengthscale of each spectral mode and thus precision will be n-dependant.

To establish a baseline for numerical precision we initially consider a range of uniform precision reductions in spectral-space. For the grid-point calculations and spectral transforms the precision is kept fixed at single precision. Precision is varied using software emulation, for speed and ease of application [Dawson and Düben, 2016]. This does not provide the computational gains of using graphics processing units or field-programmable gate arrays to implement the precision reduction but provides a flexible test setup to study the impact of a reduction in precision. As our preliminary investigation we study the evolution of hurricane Sandy. Extreme weather events are a vital test-case of any model development. Hurricane Sandy is an interesting event to consider as it presented prediction busts for many global models. While IFS correctly predicted the coast-ward turn of the hurricane, many other models predicted a continuation parallel to the coast. It is therefore a good test of precision reduction to investigate if the true hurricane behaviour is maintained.

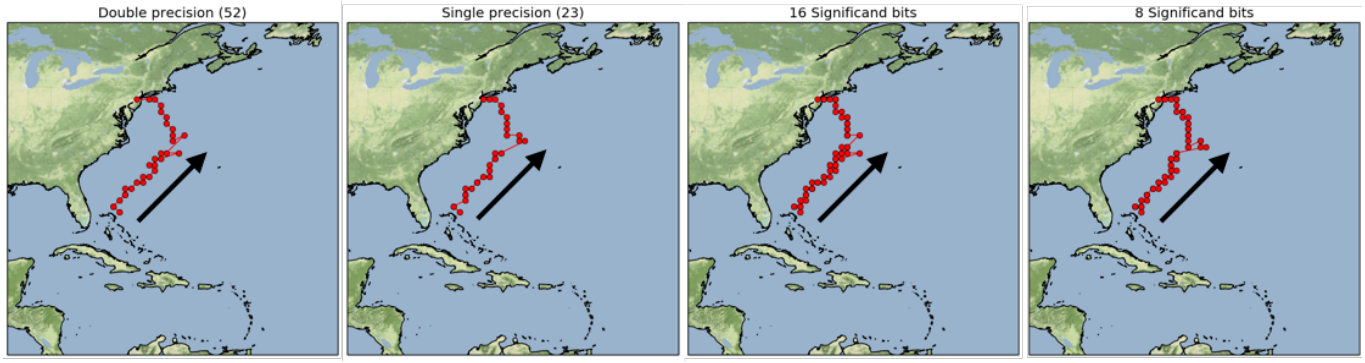

Figure 1: Evolution of hurricane Sandy at 80km resolution using OpenIFS. 3-hourly vorticity maxima at pressure level 925hP are plotted in red for simulations using 4 different precision levels in spectral space. From left to right, double precision, single precision, 16-bit significand and 8-bit significand. Although the precise position of the voriticity maxima varies between all four evolutions, the hurricane landfall position and hour is reproduced despite the heavy precision truncation.

In figure 1 we plot the 3-hourly local vorticity maxima at pressure level 925hP during a 90 hour simulation for four different precision levels. We compare double and single precisions, which use 52 and 23 bits to represent the significand, with runs using 16 and 8 bit significands. This reduction is carried out only in spectral space as this is the current scope of our investigation. In all four simulations the location and intensity of the maxima is very similar during the entire forecast. In addition, the time and location of the hurricane landfall is accurately predicted. This is highly encouraging for our more detailed investigation, and suggests that current precision levels are unnecessary. In figure 2 we plot the global geopotential height of pressure level 500hP. Here, when examining features on a global scale, we see clear differences between the 8-bit significand simulation and those at higher levels. This difference is

dominated by a global positive bias in the Z500hP level. To combat this fault, we introduce the first step of scale-selectivity. Calculations for the zeroth wavenumber in spectral space, corresponding to global mean values, are kept at single precision. All other wavenumbers are simulated with 8-bit precision as before. With this simple, and computationally inexpensive change we see a large improvement in the Z500hP field. To demonstrate this, we plot the differences between single and double precision after 120h alongside the difference between 8 significand bits (with high precision zero mode) and double precision. While there is an increase in the deviation from the double precision forecast, this deviation is small compared to the variance of the field itself. From hereon we shall use high precision zero modes for all reduced precision calculations.

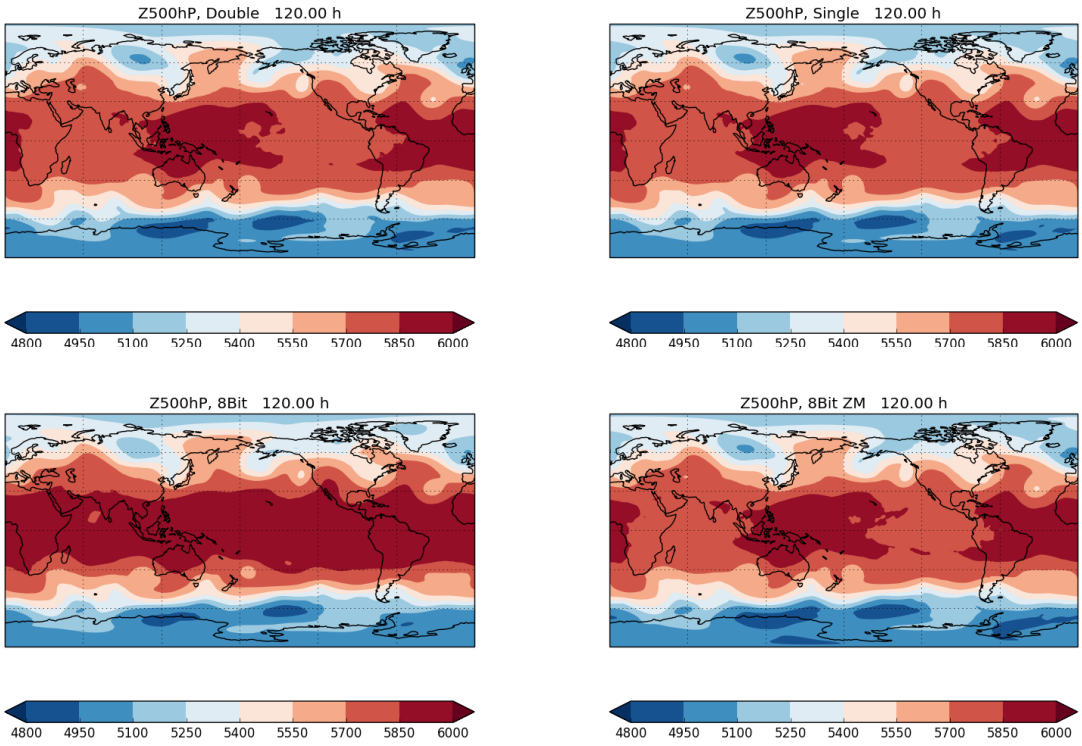

Figure 2: Geopotential height of pressure level 500hP after 120 simulated hours. Results for double and single are indistinguishable, whereas at 8-bit variations can be seen globally, dominated by a global bias. By retaining single precision for the 0th total wavenumber (global mean) this error is largely removed.

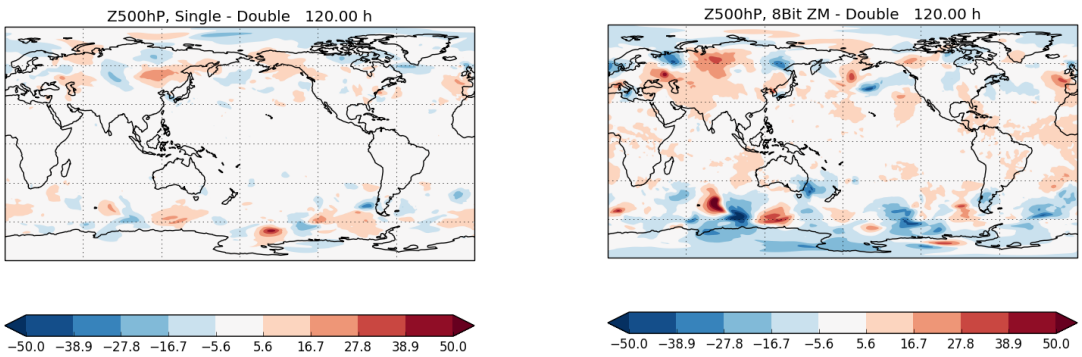

Figure 3: Deviation from double precision simulation for geopotential height of pressure level 500hP after 120 simulated hours. Using the high precision 0th wavenumber with otherwise 8-bit calculations we see only a modest increase in error when compared to single precision deviation.

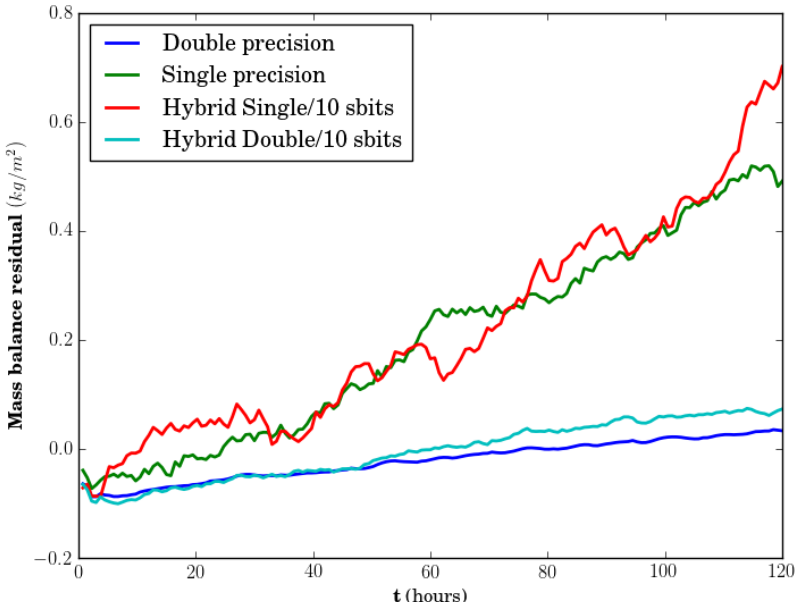

Figure 4: Mass deviation from initial condition as a function of time for four different precision levels: double precision, single precision, hybrid double precision - 10-bit significand precision in spectral space and hybrid single precision - 10-bit significand precision in spectral space. The precision level in physical space dominates the mass conservation properties. Mass is well conserved if the zeroth total wavenumber is kept at high precision.

The work introducing a single precision version of IFS exacerbates a long-standing issue of mass-conservation in the model [Váňa et al., 2017], which leads to skill degradation in the tropics. This issue is currently under investigation at ECMWF. Mass conservation is both a point for concern in any further precision reduction and an opportunity to use our variable precision version of IFS to identify where precision is causing mass loss. In figure 4 we plot the mass balance residual, the deviation in the mass from the initial mass. Firstly, we can confirm the increased mass loss in single precision OpenIFS compared to double precision. Alongside these results we carry out two variable precision experiments. The first (red) uses 10-bit significands for all wavenumbers greater than zero and single precision for the remainder of the calculations (e.g. physical space calculations and spectral transforms). The second (light blue) again used 10-bit significands for all wavenumbers greater than zero but with double precision for the remainder of the calculations. Both variable precision experiments follow closely the mass-conservation properties of their physical space precision partners. From this we can learn two related facts. Firstly, the single precision mass loss does not originate in spectral space. Secondly, even at low precision levels in spectral space mass is well conserved. These results are both useful to our ongoing research into reduced precision and to the work investigating single precision IFS at ECMWF.

The results of reduced precision in spectral space already show promising results, even when fixed precision levels are used. We now investigate how the necessary precision level changes as a function of total wavenumber. To do this we must choose a measure of the forecast quality. For this preliminary investigation we examine the spectral norm error. This measure compares the average (across wavenumbers) spectral norm as a function of model level between reference run (here double precision) with a simulated output. The output is the largest relative deviation across all model levels and 5 fields (vorticity, divergence, temperature, humidity and kinetic energy). We choose a 1% maximum error as our threshold for simulation after a one day time integration. Experiments are run with a two-tier precision set-up, for a given total wavenumber, n, wavenumbers less than n are calculated with 16-bit significands and all wavenumbers greater or equal to n are run with a prescribed significand size. In figure 4 we show the minimum precision that satisfies our 1% threshold as a function of wavenumber, n. As we increase n and restrict the low precision to higher wavenumbers the required precision decreases. For T159, a precision level greater or equal to 10-bits in the significand is required for wavenumbers less than 40. Beyond this we observe a rapid decrease in required precision. For T255 this threshold, and the rapid decrease, are delayed until n = 80. To highlight this delay we replot the minimum precision against the ratio of total wavenumber to truncation level. The collapse of the data suggests that horizontal diffusion is enabling reduced precision. Horizontal diffusion on a variable X to a diffused value $\bar{X}$ acts as

$$\bar{X}_n^m = \left\{ 1 + \frac{1}{3} \left(\frac{n(n+1)}{N(N+1)} \right)^2 \right\}^{-1} X_n^m .$$

For larger truncations, N, diffusion has an equal effect at a proportionally larger n. At operational resolutions this decrease in required precision will likely be delayed to higher wavenumbers. For both resolutions half of all total wavenumbers can be integrated with 8-bit significands or less. This corresponds to 75% of all spectral coefficients.

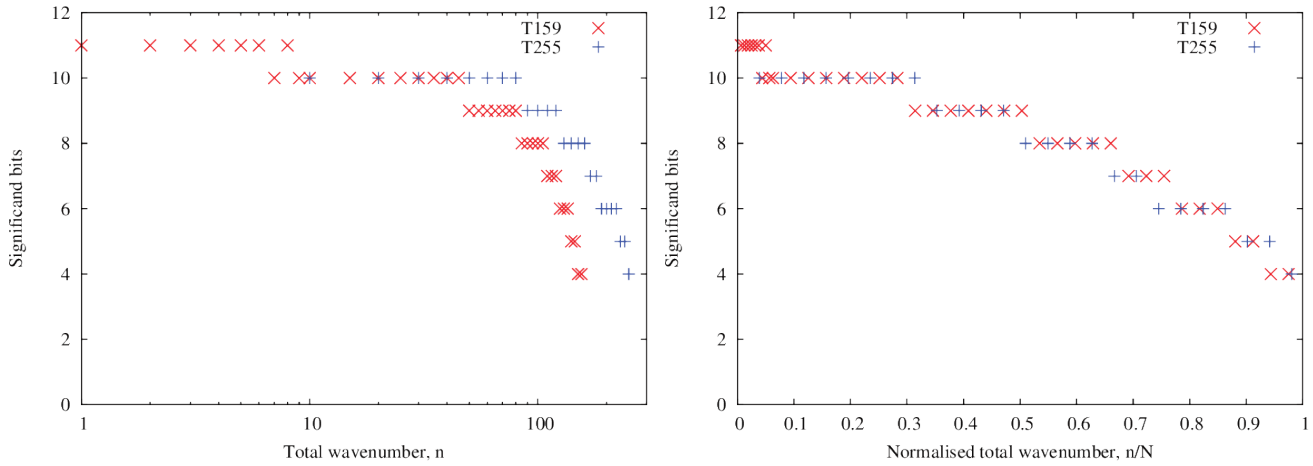

Figure 5: Minimal precision necessary to satisfy a one day 1% threshold in the spectral norm error as a function of threshold wavenumber for simulations at T159 and T255. For total wavenumbers greater than or equal to n the minimum precision required is shown on the y-axis. On the right, n is normalised by the truncation level which collapses the data.

The preliminary results suggest that 8-bit significands could be sufficient for a large portion of the spectral-space calculations. This is less than the 10-bits provided by the IEEE half-precision standard. However in our application dynamic range is of some importance, therefore redistribution of resources to create a floating point number with an 8-bit significand and a 7-bit exponent (only 1-bit less than single precision) could be the optimal choice. Currently, this level of flexibility is only available on field-programmable gate arrays. To use the newly available half-precision on GPUs would require extra thought to condense the dynamic range.

A second study to apply scale selective numerical precision was investigating a reduction in precision in the cloud resolving model of the superparametrised version of OpenIFS. Following a similar approach compared to the tests above for precision at different spacial scales in OpenIFS, we have automated the search for the ideal precision level for most model parameters in the cloud resolving model. 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 [Düben et al. 2017].

We have also implemented and tested an ensemble multi-scale modeling approach with the ECMWF IFS forecasting system for the first time in producing medium range and subseasonal-to-seasonal time range forecasts. We compare probabilistic forecasts with stochastic parameterization vs ensemble superparameterization for tropical convective systems to study the nature of convective error growth. We show that a multiscale ensemble modeling approach helps improve forecasts of certain aspects of tropical

convection during the MJO events, while it also tends to deteriorate certain largescale dynamic fields with respect to stochastically perturbed physical tendencies approach that is used operationally at ECMWF. We used units from the spgbtpia account for some of these experiments. We have published our first set of results in Subramanian and Palmer [2017]. We have further run experiments with 20 years of subseasonal hindcasts with this ensemble multi-scale modeling approach to compare to the stochastic parameterization approach. We are currently in the process of writing up these results. All these experiments were run with single precision CRM for the superparameterization experiments.

A. Dawson and P. D. Düben. rpe v5: An emulator for reduced floating-point precision in large numerical simulations. Geoscientific Model Development Discussions, pages 1–16, 2016.

P. D. Düben and T. N. Palmer. Benchmark tests for numerical weather forecasts on inexact hardware. Monthly Weather Review, 142(10):3809–3829, 2014.

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

P. D. Düben, A. Subramanian, A. Dawson, and T. Palmer. A study of reduced numerical precision to make superparameterization more competitive using a hardware emulator in the openifs model. Journal of Advances in Modeling Earth Systems, 9(1):566–584, 2017.

A. C. Subramanian and T. N. Palmer. Ensemble superparameterization versus stochastic parameterization: A comparison of model uncertainty representation in tropical weather prediction. Journal of Advances in Modeling Earth Systems.

T. Thornes, P. D. Düben, and T. N. Palmer. On the use of scale-dependent precision in earth system modelling. Quarterly Journal of the Royal Meteorological Society, 143(703):897–908, 2017.

F. Váňa, P. D. Düben, S. Lang, T. N. Palmer, M. Leutbecher, D. Salmond, and G. Carver. Single precision in weather forecasting models: An evaluation with the ifs. Monthly Weather Review, 145(2):495–502, 2017.

N. P. Wedi, M. Hamrud, and G. Mozdzynski. A fast spherical harmonics transform for global nwp and climate models. Monthly Weather Review, 141(10):3450–3461, 2013.

List of publications/reports from the project with complete references

A. C. Subramanian and T. N. Palmer. Ensemble superparameterization versus stochastic parameterization: A comparison of model uncertainty representation in tropical weather prediction. Journal of Advances in Modeling Earth Systems.

P. D. Düben, A. Subramanian, A. Dawson, and T. Palmer. A study of reduced numerical precision to make superparameterization more competitive using a hardware emulator in the OpenIFS model. Journal of Advances in Modeling Earth Systems, 9(1):566–584, 2017.

Summary of plans for the continuation of the project

(10 lines max)

The major goal of this project is to test the validity of reduced precision in a full weather model, in this case OpenIFS. The necessary coding and testing part of our reduced precision application now been completed, as highlighted above. We will now use our special project allowance to investigate reduced precision OIFS at higher resolutions, closer to the operational threshold. Specifically we will run ensemble forecasts at T511 resolution. This work will start before the end of the year and consume the rest of our remaining budget CPU time. Our group will hire several new postdoctoral researchers on the subject of reduced precision. They will investigate reduced precision in different parts of OIFS, including the spectral transform routines. These routines are expensive at operational resolutions and have bad scaling characteristics for further resolution upgrades. Reduced numerical precision could be key in reducing this numerical burden.