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	2018
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	18064420	15000000	1290257
Data storage capacity	(Gbytes)	15000	Low	15000	Low

Summary of project objectives

(10 lines max)

Investigate the possible benefits of reduced or variable numerical precision on weather and climate prediction. This builds on the work carried out in Oxford investigating reduced precision in simple models. Double precision is used as the default precision level in most weather and climate codes, yet these codes contain large sources of uncertainty and error. We are investigating if this high precision is necessary for an accurate forecast. We will examine a variety of kernels of the weather forecasting process to assess the viability of reduced-precision for operational weather forecasting.

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

In the first year of this special project we carried out preliminary investigations as to the importance of numerical precision at different horizontal lengthscales. This was achieved by introducing a reduced precision emulator into the code base of the Open IFS code, more precisely the section of code that solves the equations in spectral-space. This year we have built upon this work and moved towards operational forecasting resolutions. Our results, which can be found in the attached paper draft, examine the necessary precision to satisfy two measures of acceptable forecast quality. The first considers the L2-norm of the error away from double-precision, the second looks at the error relative to the spread of the operational ensemble. For either measure, precision commensurate with half-precision (the lowest available standard of precision) is acceptable for the vast majority of calculations. Only the largest lengthscales, approximately the first 10 modes, require higher precision. This result is found consistently across resolutions of TL159, TL255, TL511 and the scaling with resolution suggests this will also hold true for operational resolutions. This marks the end of this part of the project, which will culminate with the submission of the attached paper for peer-review.

The special project units have also been used to studying the impact of reduced precision within other areas of the Open IFS code. We are currently examining precision in the context of the physical parameterisation schemes and the Legendre transforms. Many of the physical parameterisation schemes are approximations necessitated by the limitations of computational resources or numerical resolution. Therefore calculating these schemes with low precision should be an achievable goal. The Legendre transforms are one of the most expensive areas of the current operational model, and show poor scaling for future resolution increases. The success of reduced precision in spectral-space suggests double-precision is not a necessity for the transforms, particularly given that these occur just before and after the spectral-space calculations. The preliminary results for both sections of code are extremely promising, with the use of half-precision possible for the vast majority of calculations.

List of publications/reports from the project with complete references

Chantry M, Thornes T, Duben P and Palmer T — Scale-selective precision for weather and climate forecasting (To be submitted)

Summary of plans for the continuation of the project

(10 lines max)

The work thus far examining the physics schemes and Legendre transforms has been at low resolutions, predominately TL159. The remainder of this year's units will be used to expand this work into a thorough study at higher resolutions and examining many start dates. We have also been working with ECMWF to introduce some of these changes into the IFS code. The major advantage of this will be to confirm our results in the context of ensemble forecasting and verifying using the skill measures preferred by ECMWF. The units will also be used to achieve this goal.

Scale-selective precision for weather and climate forecasting

Matthew Chantry[1], Tobias Thornes[1], Peter Düben[2], Tim Palmer[1]

[1]Atmospheric, Oceanic and Planetary Physics, University of Oxford, UK
[2]European Centre for Medium-Range Weather Forecasts, Reading, United Kingdom

June 29, 2018

Abstract

Attempts to include the vast range of lengthscales and physical processes at play in the Earth's atmosphere push weather and climate forecasters to build and more efficiently utilise some of the most powerful computers in the world. One possible avenue for increased efficiency is in using less precise numerical representations of numbers. If computing resources saved can be reinvested in other ways (e.g. increased resolution or ensemble size) a reduction in precision can lead to an increase in forecast accuracy. Here we examine reduced numerical precision in the context of ECMWF's OpenIFS model. We posit that less numerical precision is required when solving the dynamical equations for shorter lengthscales while retaining accuracy of the simulation. In spectral space calculations such a lengthscale decomposition is carried out so we introduce a reduced precision emulator into these calculations and optimise the precision necessary to achieve forecasts comparable with double and single-precision. For weather forecasting timescales a clear scale-dependance of necessary numerical precision is found. Over decadal timescales, less scale-selectivity is observed, but half-precision is still sufficient precision for the majority of calculations.

1 Introduction

Improving the efficiency of weather and climate forecasts is essential in a world struggling to predict and adapt to anthropogenic climatic change and its associated weather extremes. More accurate forecasts will require numerical models of Earth's atmosphere to become more comprehensive in the processes they capture, to operate at higher resolutions and to employ larger ensembles during simulations. All of this comes at an increased computational cost, however, and there is a limit to the computational cost of forecasts that forecasting centres can afford. While the peak performance of supercomputers continues to increase, individual processors have ceased to get faster in recent years, so that more processors must be run in parallel. However, an increase in the number of processors incurs a proportional energy cost that may become unaffordable in the near future. Using computer resources more efficiently is therefore a requirement to achieve a substantial increase in forecast accuracy. Typically simulations run at less than 5% of peak supercomputer performance as data movement, both within and between cores, dominates the computational burden.

Conventionally, most numerical models use double-precision floating-point values to represent real numbers during calculations, assigning 64 bits of precision to each number. However, there is no a priori reason why 64 bits should be the optimal number to use. Reducing the number of bits to less than the double-precision default poses the risk of increased rounding errors. For instance, single-precision (32 bits total) has a maximum relative error of 2^{-23} and can represent numbers up to approximately 3.4×10^{38}, whereas half-precision (16 bits total) represents numbers up to 65504 with a maximum relative error of 2^{-10} (IEEE 2008). There is no IEEE standard for lower precision than this, for which applications have been historically few.

No model of the weather or climate yields perfect forecasts, regardless of the precision used, as a result of observational and modelling errors. For example, the observations used to generate initial conditions for forecasts are of an inevitably limited accuracy and spatial density, and are certainly not accurate to many decimal places that would justify a representation in double-precision. Data assimilation routines are imperfect when feeding observations into the model simulation, and any numerical method to solve the non-linear equations that describe the fundamental physics of the atmosphere in the model itself will always introduce further errors. Primarily, such errors arise from the model's finite resolution, which means that there are always sub-grid-scale phenomena that cannot be explicitly resolved and whose effects must instead be approximated through parameterisation.

A number of recent studies have shown that accurate forecasts could be produced at a much lower computational cost by running parts of both full-complexity and more idealised weather and climate models in reduced precision. Düben and Palmer [2014] first demonstrated that the OpenIFS could be run entirely in single-precision (32 bits) producing reasonable results, and Váňa et al. [2017] demonstrated for IFS that model quality in single-precision was very comparable to double-precision simulations while a forty per cent reduction of run-time could be achieved. Nakano et al. [2018] showed small errors and a 46% reduction in runtime when running the Nonhydrostatic Icosahedral Grid Atmospheric Model (NICAM) with the majority of calculations at single-precision. Even lower precision has been successfully applied to the Intermediate Global Climate Model (IGCM), for which Düben and Palmer [2014] estimate that 98% of the model could be run with 20 bits to represent real numbers (8 bit significand) to yield better results than a low-resolution alternative with similar computational costs.

In the absence of real reduced-precision hardware capable of running full-complexity models in less than single-precision the effects of reduced precision can be emulated on double-precision hardware (see section 3). However, there is tangible evidence for the potential benefit of reduced precision computing. Russell et al. [2015] used Field Programmable Gate Arrays (FPGAs), which are difficult to program but can apply user-specified precision, to run the idealised Lorenz '96 model atmosphere in less than half-precision (16 bits) and demonstrated considerable speed-ups relative to double-precision with negligible loss of accuracy. With the advent of the "Volta" Graphical Processing Unit (GPU) and Tensor Core produced by NVidia half-precision hardware will become available to forecast centres (provided that forecast models are able to run on GPUs). Volta can run in double-, single- or half-precision with a linear decrease in power requirements per calculation as the precision is reduced (NVidia 2018). The Tensor Core on NVIDIA Volta GPUs is 16 times faster when multiplying half-precision matrices when compared with double-precision. There is a strong demand for 'mixed precision' architectures in 'deep learning' artificial intelligence applications, and weather and climate forecasters could benefit from this trend.

The Earth System shows non-linear and chaotic behavior. It is difficult to identify the optimal level of precision in such a system. However, to guide a reduction in precision it has been suggested to reduce numerical precision for computations of small spatial scales while keeping precision high for computations that calculate large-scale behaviour. Small scale dynamics are inherently uncertain due to the strong influence of parametrisation schemes at these scales as well as fast error growth and limited skill in the assimilation of atmospheric observations. The approach to reduce precision with spatial scale has already been demonstrated to yield accurate forecasts in models of low and medium complexity, namely Lorenz'96 and the Surface Quasi-Geostrophic (SQG) equations [Thornes et al., 2017, Thornes et al.]. It was shown that the smallest scales (highest wavenumbers) of a simulation in SQG could be represented with just 5 bits in the significand with negligible impact on the accuracy. Results in Düben and Palmer [2014] suggests that the same approach can also be realised in a three-dimensional spectral dynamical core.

This paper will describe a series of experiments designed to test whether precision can be reduced beyond single precision without impairing the forecast accuracy in the Open Integrated Forecasting System (OpenIFS) developed by the European Centre for Medium-Range Weather Forecasts (ECMWF). OpenIFS is the portable version of ECMWF's operational weather forecast model IFS. This study emulates low precision in the ECMWF OpenIFS and applies, for the first time, a scale-selective approach to this system. Precision is reduced more at the less certain small spatial scales within the spectral part of the model, where variables are represented through different wavenumber components corresponding to different spatial scales and scale-selectivity is hence possible.

The rest of this paper is organised as follows. Section 2 outlines the OpenIFS in more detail, whilst section 3 describes the implementation of emulated reduced precision therein. Section 4 describes a number of experiments designed to test the proficiency of scale-selective precision in OpenIFS and presents the results. The findings are discussed in section 5, and the paper concludes with a discussion of the potential implications.

2 OpenIFS

The European Centre for Medium-Range Weather Forecasts produces global forecasts using the Integrated Forecast System (IFS). At the time of writing, the latest operational model is version 43r3, released in July 2017, which produces a single ten-day global forecast at 9 km horizontal resolution with 137 vertical levels and a fifteen-day fifty-member ensemble global forecast at 18 km horizontal resolution with 91 vertical levels (ECMWF 2017a).

For research activities external to ECMWF the centre makes available a portable version of the model, called "OpenIFS", which is available for licensed researchers to download and use remotely. The OpenIFS has all the forecast functionality of the operational IFS including all the parameterisation schemes, and contains some half a million lines of code distributed across more than two thousand files [ope]. The version of OpenIFS that is employed here is based on the '38r1' release of IFS, which was used operationally until June 2013, in a fully-functional format except for the data assimilation and ensemble-forecasting components. In this study the maximum resolution at which OpenIFS is run is 'TL511', which corresponds to 512 wavenumber components (from 0 to 511) in spectral space and a triangular-linear reduced-Gaussian grid of approximately 40 km horizontal resolution. By contrast, from January 2010 to March 2016 (a timespan covering most of the test-cases presented here) the ECMWF's operational releases of IFS used a "TL639" resolution for its ensemble, which corresponds to 640 wavenumber components and roughly 32km horizontal resolution.

The OpenIFS uses a semi-Lagrangian, semi-implicit numerical scheme to solve the Navier–Stokes equations for the momentum, surface pressure, temperature, geopotential and vertical velocity of atmospheric fluid parcels at each timestep. For the implicit part of the timestepping fields are transformed to spectral space through Fourier and Legendre transforms.

3 Reduced precision emulator

Conventionally, most numerical models use double-precision floating-point values to represent all variables during calculations, assigning 64 bits of precision to each number. Of these, 1 bit represents the sign of the number (+1 or -1), E bits represent the exponent (the highest power of two that the number is greater than) and S bits represent the significand (the number's exact multiple, somewhere between 1 and 2, of this power) according to standards set by the IEEE (2008). The overall number is given by,

$$N = \pm S \times 2^{E},$$

where the value of each bit in the significand, b_i, or the exponent, c_j, is either 0 or 1, the significand and exponent are given by,

$$S = 1 + \sum_{i=0}^{51} b_i 2^{-i}, \quad E = \sum_{j=0}^{10} c_j 2^{j} - 1023.$$

This means that the significand is a fraction between 1 and 2. All "normal" numbers representable in double-precision lie between 2^{-1022} and 2^{1023}, with minimum a spacing of 2^{-41}. Any number outside this spacing will be rounded to the nearest resolvable value; numbers outside the resolvable range will be rounded to zero or infinity.

To emulate the effects of reduced precision hardware that does not strictly follow the IEEE standard requires a special 'emulator' program to be compiled and run alongside the standard modules included within the OpenIFS Dawson and Düben [2016]. Alterations to the main program that are required are minimal. The modified program induces reduced precision by treating variables as Fortran derived types that are defined by the emulator. For each operation involving these derived types the values are passed in, operated on, and then truncated to a specified number of bits. In this way, the main program always runs using double-precision floating point numbers and double-precision hardware throughout, but its output is the same as that would be obtained using mixed precision hardware. The extra costs associated with emulated precision result in computations that are slower than standard double-precision arithmetic. Real mixed precision hardware would not go through this process of rounding numbers and would therefore entail no such costs. Hence, the emulator cannot be used to analyse the potential computational cost savings that such hardware would yield, only the effect that it might have on the accuracy of the model being run.

The size of OpenIFS's code-base make the complete introduction of the emulator a challenging undertaking. When combined with the increased computation overhead for running the emulator, we decide to introduce the emulator in only a portion of the code-base. Building upon the scale-selective work of Thornes et al. we select those computations carried out in spectral space. Although this area has a relatively small computational cost **?**, there are interesting scientific questions to be asked for the information content required for these modes. The largest uncertainties and shortest predictable timescales are for short lengthscales **?**. In spectral space, fields are decomposed by lengthscale and phase, represented by complex spectral coefficients. Linear calculations on these coefficients are then carried out, with no interaction between coefficients representing sufficiently different lengthscales (derivatives of fields involve interactions between adjacent modes). This enables the use of different precision levels for calculations at different lengthscales, which we shall here investigate.

In the timestepping loop, prognostic fields and their increments are transformed from grid-point to spectral space. Horizontal wind components are converted to a vorticity and divergence representation. Beyond this point reduced precision is used for all calculations up to the point where vorticity and divergence are converted back to wind components. This choice for reduced precision introduction was motivated by the code structure. Here, global precision reduction will refer to a fixed precision level being used for all spectral space calculations. Scale-selective precision will involve precision dependent upon the total wavenumber, n. This is achieved by element-wise changes to the number of significant bits used for all vectors and arrays containing spectral coefficients. While vector and

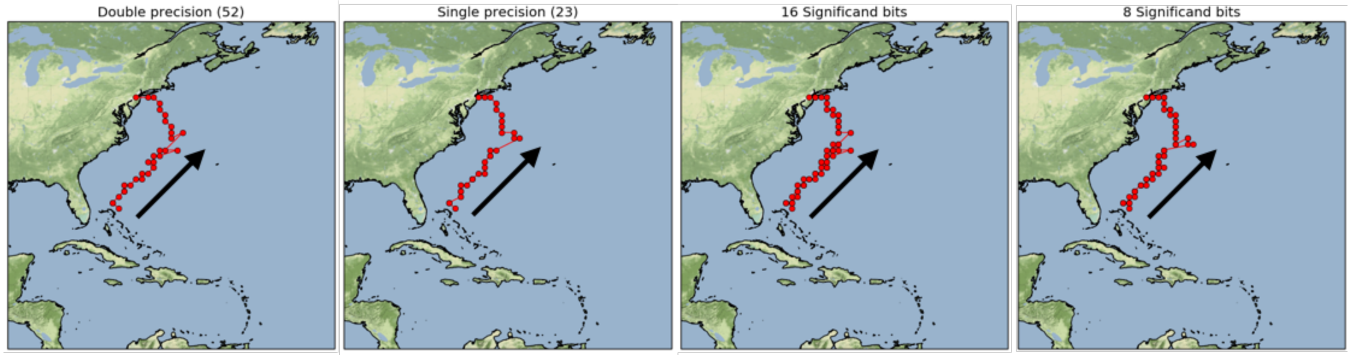

Figure 1: Evolution of hurricane Sandy at 80km resolution using OpenIFS. 3-hourly vorticity maxima at pressure level 925hP are plotted in red for simulations using 4 different precision levels in spectral space. From left to right, double-precision, single-precision, 16-bit significand and 8-bit significand.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

array calculations dominate, there are also calculations involving scalar fields and literals which take their precision from a global significant bits variable. This number has a limited impact on the accuracy of our forecasts, but is typically set to match the highest precision used by the scale-selective values. This choice, made to limit the changes to the code base, could permit precision to leak in individual calculations, but the outputs of these calculations are stored in vectors or arrays with scale-selective prescribed precision where the correct precision will be restored. On real hardware the slowdown introduced by occasional use of double-precision variables will not be significant as flointing-point operations (flops) are typically not the computational bottleneck.

4 Results

4.1 Test case: Hurricane Sandy

We begin with a reforecast of hurricane Sandy, to test the impact of global precision reduction. In figure 1 we plot the hurricane path from the forecast start-date, 2012/10/27, until landfall on 2012/10/29. A resolution of T255 (80km horizontal grid), with 91 vertical levels is used, with a time step of 45 minutes. Although the hurricane centre varies slightly between precision levels, the strength (not plotted), landfall location and landfall time are constant for precision levels down to 8 significand bits.

In contrast, the geopotential height of pressure level 500hP (Z500) is significantly changed by this lowest precision (figure 2). After 5 days, there is a clear global bias in Z500. This can be ascribed to representing large mean quantities, such as geopotential or temperature with few significand bits. The global mean for the double-precision Z500 is 5648.47m, compared to the 8-bit significand value of 5690.89m. Representing values of this magnitude with 8-bit significands can only be achieved to the nearest 16m, an unacceptable level of accuracy. This issue of large global means also affects the temperature field. Values with magnitude 300 have a spacing (between neighbouring representable numbers) of approximately 1 degree. Here, the use of Kelvin instead of Celsius may be considered a waste of bits. Rewrite 300 Kelvin as 26.8125 Celsius and the spacing is decreased below 0.1C. Changing units could, in the right context, change the viability of using half-precision in general circulation models. In spectral space, there is a clear route forward: the precision used should be dependent on the total wavenumber. The zeroth mode of a spherical harmonic expansion represents the global mean of the field, where the unit choice of fields is the most significant. In figure 2(d), we plot Z500 for a forecast with double-precision used for calculations in spectral space involving the zeroth spectral mode, and 8 significand bits for all other modes. This produces a global mean Z500 of 5649.54m, to be compared with double, 5648.47m and single, 5648.79m.

4.2 Error measures

The hurricane Sandy forecast demonstrates the value of scale-selectively setting the precision as a function of wavenumber, but lacks rigour when it comes to finding the optimal precision. To this end we introduce two measures which we will use to assess reduced precision experiments.

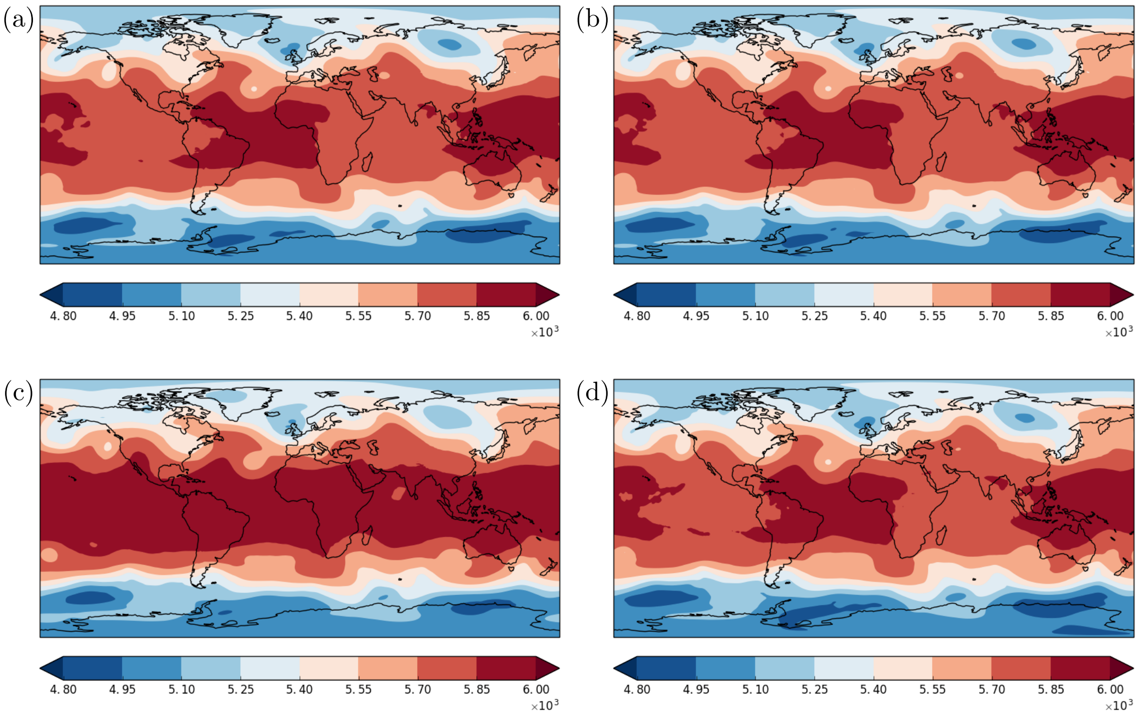

Figure 2: Geopotential height of pressure level 500hP after 120 simulated hours. Differences between (a) double and (b) single are small, whereas at 8 significant bit (c) variations can be seen globally, dominated by a global bias. (d) By retaining double-precision for the zeroth total wavenumber (which includes the global mean quantities) this error is largely removed.

The first error measure uses the distance between double-precision and single-precision (for spectral space) forecasts. Váňa et al. [2017] noted noticeable reduction in accuracy when running IFS with a majority of calculations in single-precision. We define the horizontal L_2-norm over model levels as

$$L_2^2\left(f\right) = \int_{\theta=0}^{2\pi} \int_{\lambda=0}^{\pi} f(\lambda, \theta, z)^2 \sin\lambda \; d\lambda \; d\theta.$$

We define Er as the supremum of the ratio of horizontal L_2-norm errors between a field integrated forward at reduced precision, f_r relative to the distance between double, f_d and single-precision, f_s, integrations,

$$Er(f) = \sup_z \left(\frac{L_2^2\left(f_r\left(\lambda, \theta, z\right) - f_d(\lambda, \theta, z)\right)}{L_2^2\left(f_s(\lambda, \theta, z) - f_d(\lambda, \theta, z)\right)} \right),$$

where $nlev$ are the model levels. We consider Er for prognostic variables surface pressure, temperature, vorticity and divergence at day two. A reduced precision forecast is considered acceptable if Er is less than 2 for all 4 fields. Calculating this measure after longer integration times consistently gave weaker precision constraints.

The second error measure used here attempts to use information from the ECMWF ensemble standard deviation to capture the uncertainty in the model. Considering Z500 over Europe, we calculate the proportion of gridpoints that lie more than one standard deviation away from a double-precision forecast. The European region is used because the test-cases were chosen on the basis of selecting a wide variety of atmospheric conditions in the European region. Assuming normality, we consider a forecast acceptable if less than a third of gridpoints lie more than one standard deviation of the operational forecast ensemble from the double-precision forecast. This is measured over days 2 to 5 of the forecast. Day 1 is excluded as the ensemble standard deviation is very small during day 1, so an accurate 5 day forecast appears inaccurate if precision is adjusted to day 1.

For the study below, we use resolutions significantly below operational values which prevents the fair comparison of reduced precision models using observations. Our aim with the two measures introduced is to provide approximate upper and lower bounds on acceptable precision. The L_2-norm sets a very tight error threshold, aiming to provide a model very close to double-precision. In contrast, the ensemble-spread measure attempts to create a model within the uncertainty of an ECMWF probabilistic forecast. Given that the probablistic forecast incorporates initial condition

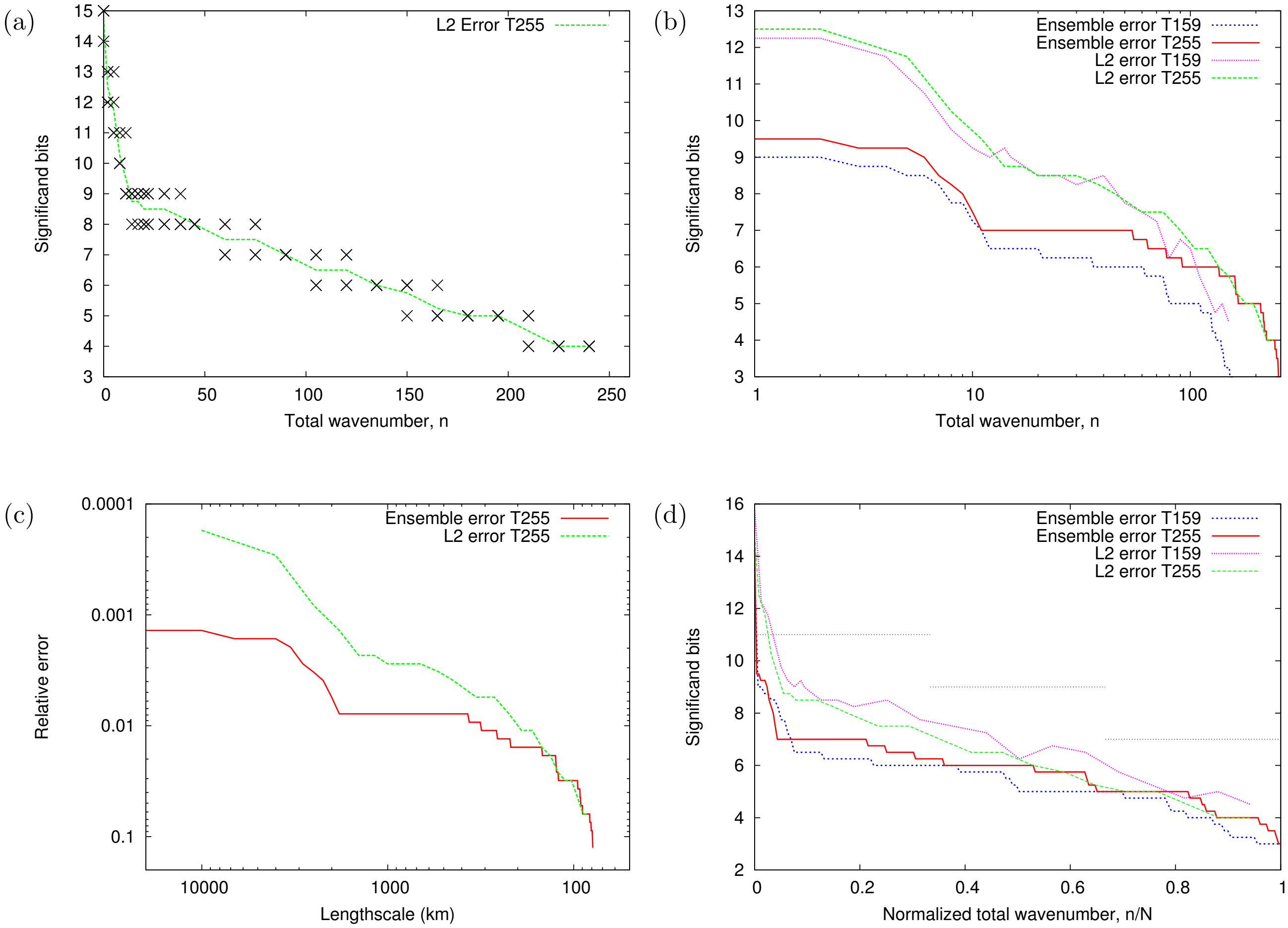

Figure 3: (a) Precision needed to represent total wavenumbers $\geq n$ for an L_2-norm error less than two. Crosses represent four start-dates at T255 resolution, with the line denoting the average. (b) Average precision for European ensemble error or L_2 error measures at resolutions T159 and T255. (c) Replotted T255 data as relative error against lengthscale. Please note that the y-axis is reversed. (d) Total wavenumber is normalized by the truncation wavenumber to collapse the data for each measure. Grey dotted lines indicate scale-selective precision used for the decadal runs (see section 4.4).

uncertainty and stochastic physics (ECMWF), neither of which are part of our reduced precision models, this creates a weaker error constraint. Together these two should provide a guide for an optimal precision setup.

We consider four start dates for the following study, 2009/10/26 00:00, 2010/02/11 00:00, 2013/10/26 00:00 and 2014/02/11 00:00. These dates cover two recent UK storm conditions in 2013 and 2014, as well as the same calendar date four years earlier which exhibited calm conditions for the UK.

4.3 Optimal precision

In figure 3(a) we plot the necessary precision for total wavenumbers greater than or equal to n to satisfy the L_2-norm measure for four T255 integrations (crosses) and the average. There is little start-date dependence and a trend towards lower precision being required at higher wavenumbers. A similar pattern is found when reducing precision only for a single total wavenumber (not plotted). Figure 3(b) plots the average precision against wavenumber for both norms at two resolutions: T159 and T255. The ensemble error gives a lower precision requirement, particularly for small wavenumbers. For both datasets, noticeably higher precision is required for the first five wavenumbers. As the truncation limit is reached, the necessary precision rapidly decreases to levels equivalent to one significant figure. This is illustrated in figure 3(c) where the T255 data is transformed to lengthscale and relative error using the formulae

$$\text{Relative error} = 2^{-\text{Significand bits}}, \;\; \text{Lengthscale} = \frac{C_{Earth}}{2n}.$$

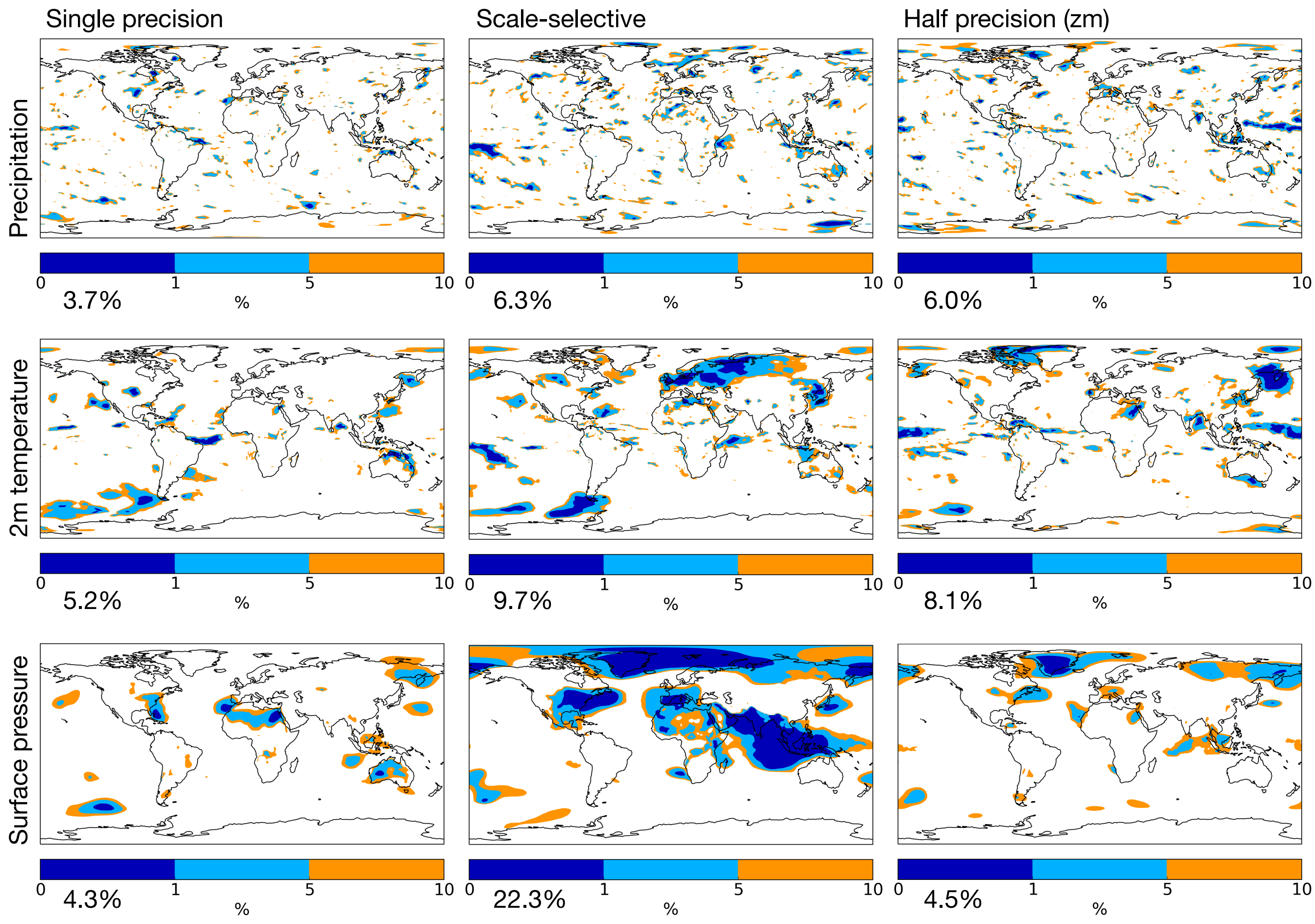

Figure 4: Significance testing 10 member climate ensembles at varying precision. Paired t-tests are carried out against a single-precision 10 member ensemble. For each field tested significance levels are plotted for 1%,5% and 10% probabilities. The proportion the globe that is significant at the 5% level is written below for accumulated precipitation, 2m temperature and surface pressure averaged over 10 years.

After rescaling the data by N, the truncation limit, we see good agreement between the two different resolutions (figure 3d). This suggests a trend whereby necessary precision is a function of normalized total wavenumber, n/N. Extrapolating to operational resolutions, this scaling would predict a delay in total wavenumber from which low precision could be used. This will required a high resolution study to investigate. The number of bits used to store the state vector could be significantly decreased if only the bits used for integration were stored. Even when maintaining a double-precision length exponent (11 bits), the number of bits used to store the T255 spectral space vector is decreased by over 70%. The observed scaling of precision as a function of n/N results in savings that are independent of the resolution.

4.4 Decadal runs

The motivations for reduced numerical precision for weather forecasts are equally valid for climate predictions. Here we wish to test the impacts of reduced numerical precision on long time integration. For this we run the following experiment: initial conditions from 1st-10th of January 2005 are integrated forward to the end of 2015 with prescribed observed SSTs. Discarding the data from 2005 we have a ten member ensemble for the decade 2006-2015. The model resolution is T159 with 91 levels and is forced with identical sea-surface temperatures for all runs. We consider three different numerical precisions: single, scale-selective and half-precision. Each is compared with a double-precision 'truth' ensemble. Scale-selective precision is crafted using the results from weather prediction. Specifically, scale-selective means double-precision for the zero mode, 11 significant bits for the first 50 total wavenumbers, 9 significand bits for the next 50 and 7 significand bits for the higher modes (see dashed grey lines in figure 3c). This is generally a cautious approach compared to our results, with the exception of the first six (non-zero) modes for which necessary

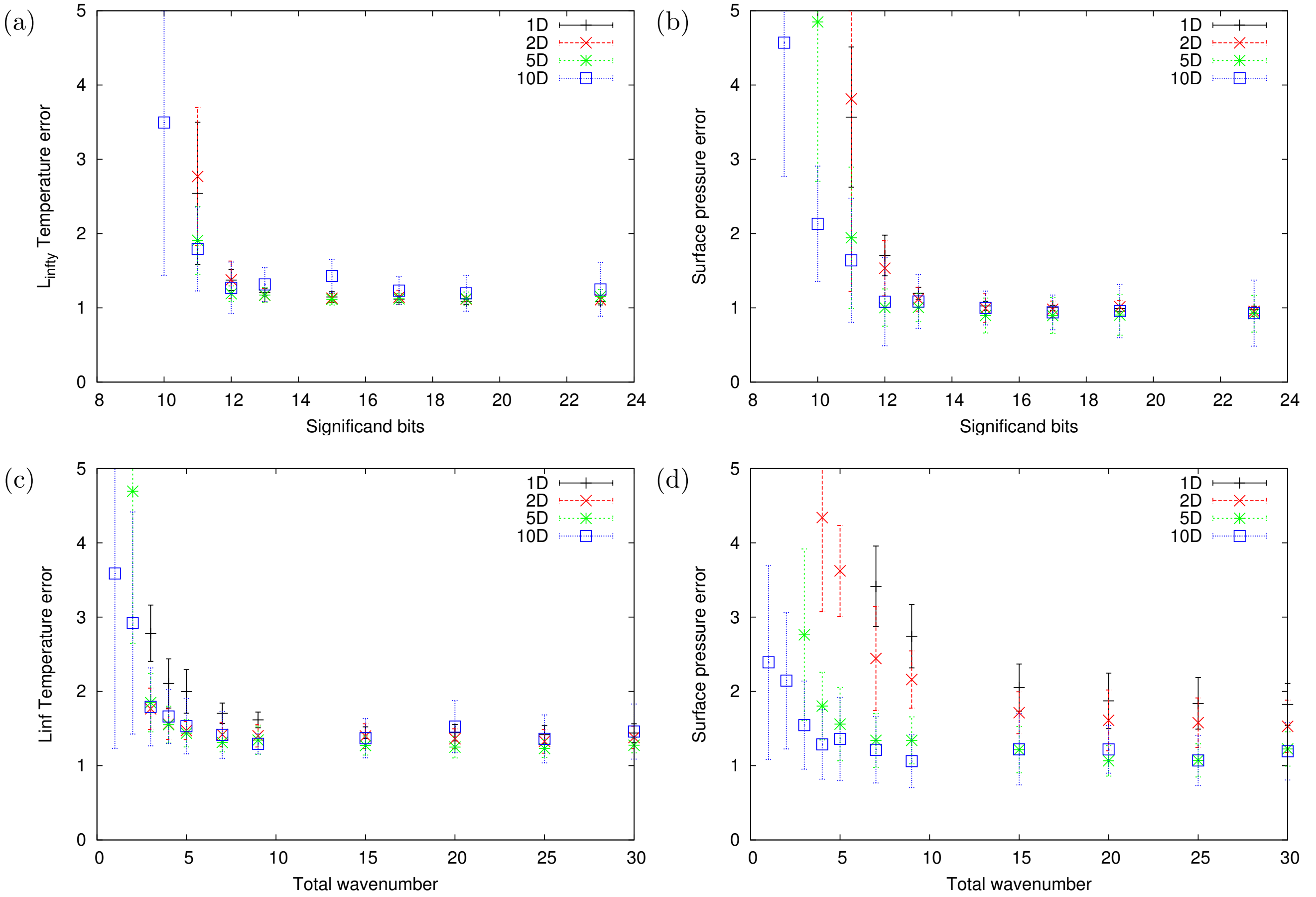

Figure 5: Reduced precision T511 experiments compared to the distance between double and single-precision, averaged over 9 start dates after 1,2,5 and 10 days of integration. (a) Temperature error for uniform precision reduction in spectral space. (b) Surface pressure error for uniform precision reduction. (c-d) Fixing global precision at 12 significand bits and using 10 significand bits from varying total wavenumber (indicated on the x-axis) for the same error measures of temperature and surface pressure. For temperature errors beyond 1 day, 10 significand bits has no increase on error for wavenumbers 7 and above. For surface pressure, global 12 significand bits introduces some difference at early times which are no longer significant at 5 days.

precision decreases with required lead time. Half-precision uses the 10 bit significand for all non-zero modes, with the zeroth mode calculated in double-precision (indicated by "zm"). The exponent for all precision levels is kept at the double-precision value of 11 bits. For each precision we calculate the paired T-test, testing for significantly different decadal averages of precipitation, 2m temperature and surface pressure from the double-precision 10 member ensemble. In figure 4, we plot maps of the probability implied from the t-test at the 1%, 5% and 10% levels. Beneath each we calculate the proportion of the globe that is significant at the 5% level for each precision experiment and field. For single-precision, this proportion is small and close to the 5% expected if grid-points were uncorrelated and from insignificantly different models. For our scale-selective model we find small changes to precipitation and 2m temperature but a large change in the surface pressure. For half-precision (zm), we find a small increase in proportion of significant area relative to single-precision but less than that of scale-selective. This result suggests that the scale-selectivity found in our weather optimisation is necessary for capturing the information content of the initial condition rather than the algorithmic error and that a different optimal precision is required for longer timescales. The small degradation for half-precision suggests that given the computational resources to optimise over climatological timescales a heavily reduced precision climate code would be effective.

4.5 Higher resolution weather forecasts

The low computational cost of running at T159 and T255 enabled a thorough search of optimal precision and a guide for the necessary precision at operational resolutions. To test the effectiveness of this guide we now consider

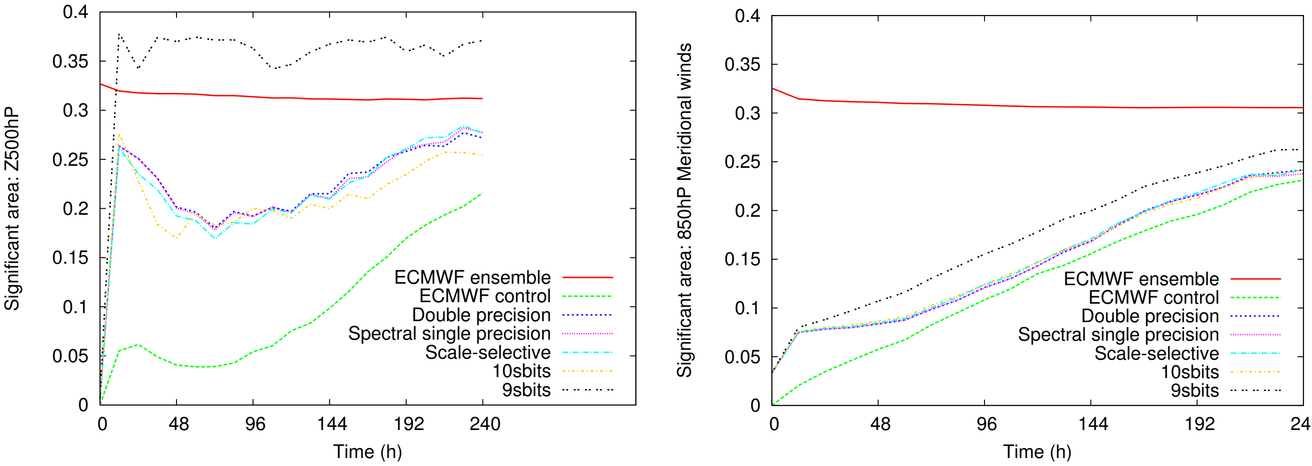

Figure 6: Proportion of globe which lies more than one ECMWF ensemble standard deviation from the ECMWF ensemble mean for (a) Z500 and (b) 850hP meridional winds. Members of the ECMWF ensemble remain close to a third, while the unperturbed ECMWF control climbs from zero. Our double-precision, single-precision, scale-selective and global 10 significand bits runs lie between these two, with precision choices having little influence until global precision is reduced to 9 significand bits. Model version and resolution changes are responsible for the difference between the ECMWF control and double-precision.

higher resolution forecasts at reduced precision. We consider 9 start dates between 2011 and 2017 at a resolution of TL511 with 91 levels using a timestep of 15 minutes. These start dates cover a range of months and conditions. The start dates considered are 2011/04/08 00:00, 2012/10/27 00:00, 2013/02/01 00:00, 2014/03/15 00:00, 2015/10/02 12:00, 2016/03/07 00:00, 2016/06/30 00:00, 2017/08/25 00:00, 2017/09/05 12:00. The effective grid resolution of approximately 39km is closer to ECMWF's 20km resolution (recently upgraded from 31km in 2016). Using the same L_2-norm as before we plot the average error across start dates for temperature and surface pressure using scale-constant precision (figure 5(a-b)). Double-precision is used for the zeroth mode. As significand bits is decreased, the error remains comparable to single-precision (close to 1) until 12 significant bits, below this the error increases rapidly. For surface pressure we see a small increase in error for 12 significand bits at 1 and 2 days which has no obvious impact on the 5 and 10 day values. Motivated by the fact that already available hardware supports 10 significand bits (half-precision) we next ask from which total wavenumber can we use 10 significand bits. In figure 5(c-d) we plot the temperature and surface pressure errors for simulations which use 12 significand for wavenumbers between 1 and $n-1$ and 10 significand bits for n and greater. Beyond $n = 15$ we see no further impact of using 10 bits on our results (relative to using 12 bits globally). Considering only the error at days 5 and 10 this can be decreased to the 5th wavenumber. A model with 10 significand bits from wavenumber 5 onwards treats over 99.98% of spectral coefficients with half-precision significands.

Finally we assess the high resolution simulations in the context of the ECMWF operational ensemble. For each start date and precision level we use the operational ensemble mean and ensemble standard deviation and calculate the proportion of the globe that lies more than one standard deviation away from the mean. This measure is plotted against time for Z500 and 850hP meridional winds for a selection of precision experiments. Here scale-selective is defined as 12 significand bits for wavenumbers less than 5 and 10 significand bits for higher wavenumbers. Model version and resolution differences account for differences between the ECMWF control (green) and our double-precision simulations (dotted blue), which otherwise use the same unperturbed initial condition. Relative to this difference, any precisions with more than 9 significand bits have equivalent performance. Perturbed members of the ECMWF ensemble lie close to the value of one-third. This would be expected randomly drawn data from a normal distribution with the ensemble mean and ensemble standard deviation. These results are highly promising for doing operational forecasting at reduced precision. Relative to model changes and initial condition uncertainty reduced precision calculations have a small effect on this set of calculations.

5 Conclusion

We have presented here the effects of reduced precision in spectral space on the accuracy of ECMWF's OpenIFS system. This marks the latest step in the assessment of reduced precision for a hierarchy of weather and climate models. The necessary precision for an accurate weather model has clear scale-separation, with large spatial scales requiring higher precision than small scales. For decadal runs, the necessary precision appears to be much lower than the double-precision currently used, but using the scale-selective precision derived from 5-day forecasts gives worse results than half-precision (with the zeroth-mode mode retained at high-precision). Computational constraints have prevented a full search for optimal precision over this timescale. This would be required to establish if the scale-selective approach is not appropriate for climatological runs, or if it merely differs from the optimal for short timescales. Towards operational resolutions we continue to find that double-precision is unnecessary, particularly when examined in the context of ensemble forecasting. When compared to the ECMWF ensemble spread, precision errors are small, but future work will need to study the spread of an ensemble when integrated at reduced-precision.

This work can also act as a first guide for the precision necessary for other calculations involved in a general circulation model. For example the Legendre transforms exhibit expensive scaling properties as resolution is increased, due to the matrix-multiplications involved. Recent hardware has shown a greater than linear speed-up when computing these at half-precision instead of double-precision. Our work here suggests that half-precision is plausible for these calculations. Investigating the spectral transforms will be carried out in future work.

References

Open IFS. `https://software.ecmwf.int/wiki/display/OIFS/OpenIFS+Home`. Accessed: 2018-04-30.

A. Dawson and P. D. Düben. rpe v5: An emulator for reduced floating-point precision in large numerical simulations. *Geosci. Model Dev. Discuss., in review, doi*, 10, 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.

ECMWF. Ifs documentation ? cy43r3: Part v: Ensemble prediction system. Technical report. URL `https://www.ecmwf.int/sites/default/files/elibrary/2017/17737-part-v-ensemble-prediction-system.pdf`.

M. Nakano, H. Yashiro, C. Kodama, and H. Tomita. Single precision in the dynamical core of a nonhydrostatic global atmospheric model: Evaluation using a baroclinic wave test case. *Monthly Weather Review*, 146(2):409–416, 2018. doi: 10.1175/MWR-D-17-0257.1. URL `https://doi.org/10.1175/MWR-D-17-0257.1`.

F. P. Russell, P. D. Düben, X. Niu, W. Luk, and T. N. Palmer. Architectures and precision analysis for modelling atmospheric variables with chaotic behaviour. In *Field-Programmable Custom Computing Machines (FCCM), 2015 IEEE 23rd Annual International Symposium on*, pages 171–178. IEEE, 2015.

T. Thornes, P. D. Düben, and T. N. Palmer. Scale-selective precision in the surface quasi-geostrophic eqations.

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.