Scale-Selective Precision for Weather and Climate Forecasting

MATTHEW CHANTRY, TOBIAS THORNES, AND TIM PALMER

Atmospheric, Oceanic and Planetary Physics, University of Oxford, Oxford, United Kingdom

PETER DÜBEN

European Centre for Medium-Range Weather Forecasts, Reading, United Kingdom

(Manuscript received 28 August 2018, in final form 23 November 2018)

ABSTRACT

Attempts to include the vast range of length scales and physical processes at play in Earth's atmosphere push weather and climate forecasters to build and more efficiently utilize 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 Open Integrated Forecast System (OpenIFS) model. We posit that less numerical precision is required when solving the dynamical equations for shorter length scales while retaining accuracy of the simulation. Transformations into spectral space, as found in spectral models such as OpenIFS, enact a length scale decomposition of the prognostic fields. Utilizing this, we introduce a reduced-precision emulator into the spectral space calculations and optimize the precision necessary to achieve forecasts comparable with double and single precision. On weather forecasting time scales, larger length scales require higher numerical precision than smaller length scales. On decadal time scales, half precision is still sufficient precision for everything except the global mean quantities.

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 centers 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 or even the inability to represent the number. 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 65 504 with a maximum relative error of 2^{-10} (Markidis et al. 2018). 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 modeling errors. For example, the observations used to generate initial conditions for forecasts are of an inevitably limited accuracy and

Corresponding author: Matthew Chantry, matthew.chantry@physics.ox.ac.uk

DOI: 10.1175/MWR-D-18-0308.1

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 nonlinear 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 subgrid-scale phenomena that cannot be explicitly resolved and whose effects must instead be approximated through parameterization.

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 idealized weather and climate models in reduced precision. Düben and Palmer (2014) first demonstrated that the Open Integrated Forecast System (OpenIFS) developed by the European Centre for Medium-Range Weather Forecasts (ECMWF) 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 40% reduction of runtime 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 forecasts than a double precision but lower-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 idealized Lorenz'96 model atmosphere in less than half precision (16 bits) and demonstrated considerable speedups 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 centers (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 (Markidis et al. 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. This paper is part of a body of work investigating whether weather and climate forecasters could benefit from this trend.

The Earth system shows nonlinear 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 behavior (Düben and Palmer 2014; Thornes et al. 2018). Small-scale dynamics are inherently uncertain because of the strong influence of parameterization 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 quasigeostrophic (SQG) equations (Thornes et al. 2017, 2018). It was shown that the smallest scales (highest wavenumbers) of a simulation in SQG could be represented with just five bits in the significand with negligible impact on the accuracy. Results in Düben and Palmer (2014) suggest that the same approach can also be realized 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 ECMWF OpenIFS. 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 organized as follows. Section 2 outlines the OpenIFS in more detail, while 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

ECMWF produces global forecasts using the IFS. At the time of writing, the latest operational model is version 43r3, released in July 2017, which produces a single 10-day global forecast at 9-km horizontal resolution with 137 vertical levels and a 15-day, 50-member ensemble global forecast at 18-km horizontal resolution with 91 vertical levels (ECMWF 2017).

For research activities external to ECMWF the center 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 parameterization schemes, and contains some half a million lines of code distributed across more than 2000 files. 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 time span 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 32-km 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 time step. For the implicit part of the time stepping, fields are transformed to spectral space through Fourier and Legendre transforms.

3. Reduced-precision emulation in OpenIFS

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 2 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 (Zuras et al. 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

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

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 2017). 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. Each variable of this type contains a value and a precision (number of significand bits). For each operation involving these derived types the values are passed in, operated on, and then truncated to the 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 analyze 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. In this study we investigate the impact of the number of bits used in the significand, which controls the precision of a floating-point number. The IEEE standard representations of double, single, and half precision use 52, 23, and 10 bits, respectively, to represent the significand.

The size of OpenIFS's code base makes 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. (2018) 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 time scales are for short length scales. In spectral space, fields are decomposed by length scale and phase, represented

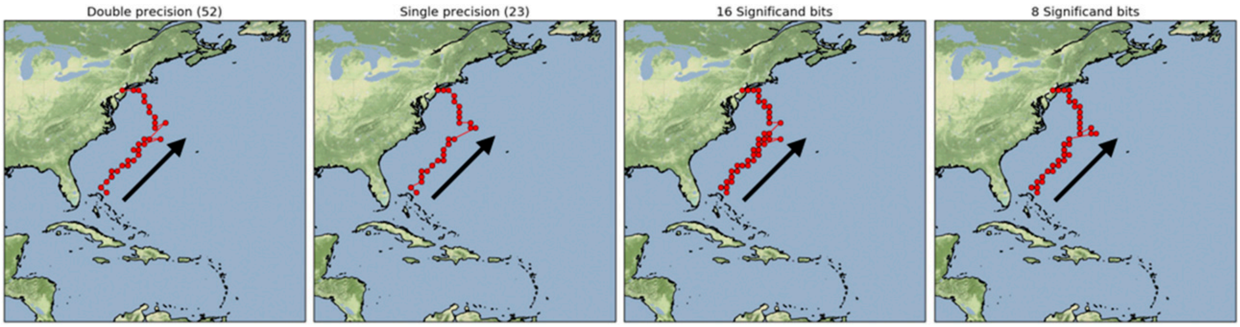

FIG. 1. Evolution of Hurricane Sandy at 80-km resolution using OpenIFS. The 3-hourly vorticity maxima at pressure level 925 hPa are plotted in red for simulations using four different precision levels in spectral space: (from left to right) double precision, single precision, 16-bit significand, and 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.

by complex spectral coefficients. Linear calculations on these coefficients are then carried out, with no interaction between coefficients representing sufficiently different length scales (derivatives of fields involve interactions between adjacent modes). This enables the use of different precision levels for calculations at different length scales, which we shall here investigate.

In the time-stepping loop, the explicit terms in the dynamics equations are transformed from gridpoint 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 fields 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 significand bits used for all vectors and arrays containing spectral coefficients.

While vector and array calculations dominate, there are also calculations involving scalar fields and literals, which take their precision from a global significand bits variable. This number has a limited impact on the accuracy of our forecasts, but is typically set to match the highest precision used in array and vector calculations. This choice, made to limit the changes to the code base, could permit precision to leak in individual calculations. However, 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 scalar variables will not be significant because floating-point operations (flops) are typically not the computational bottleneck.

4. Results

a. *Test case: Hurricane Sandy*

We begin with a hindcast of Hurricane Sandy, to test the impact of global precision reduction. In Fig. 1 we plot the hurricane path from the forecast start date, 27 October 2012, until landfall on 29 October 2012. A resolution of T255 (80-km horizontal grid), with 91 vertical levels is used, with a time step of 45 min. Although the hurricane center 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 500 hPa (Z500) is significantly changed by this lowest precision (Fig. 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 5648.47 m, compared to the 8-bit significand value of 5690.89 m. Representing values of this magnitude with 8-bit significands can only be achieved to the nearest 16 m, an unacceptable level of accuracy. This issue of large global means also affects the temperature field. Values with magnitude 300 have a spacing (between neighboring representable numbers) of approximately 1°. Here, the use of kelvin instead of Celsius may be considered a waste of bits. Rewrite 300 K as 26.8125°C and the spacing is decreased below 0.1°C. Renormalizing the variables, 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 Fig. 2d, we plot Z500 for a forecast with double precision used for calculations in

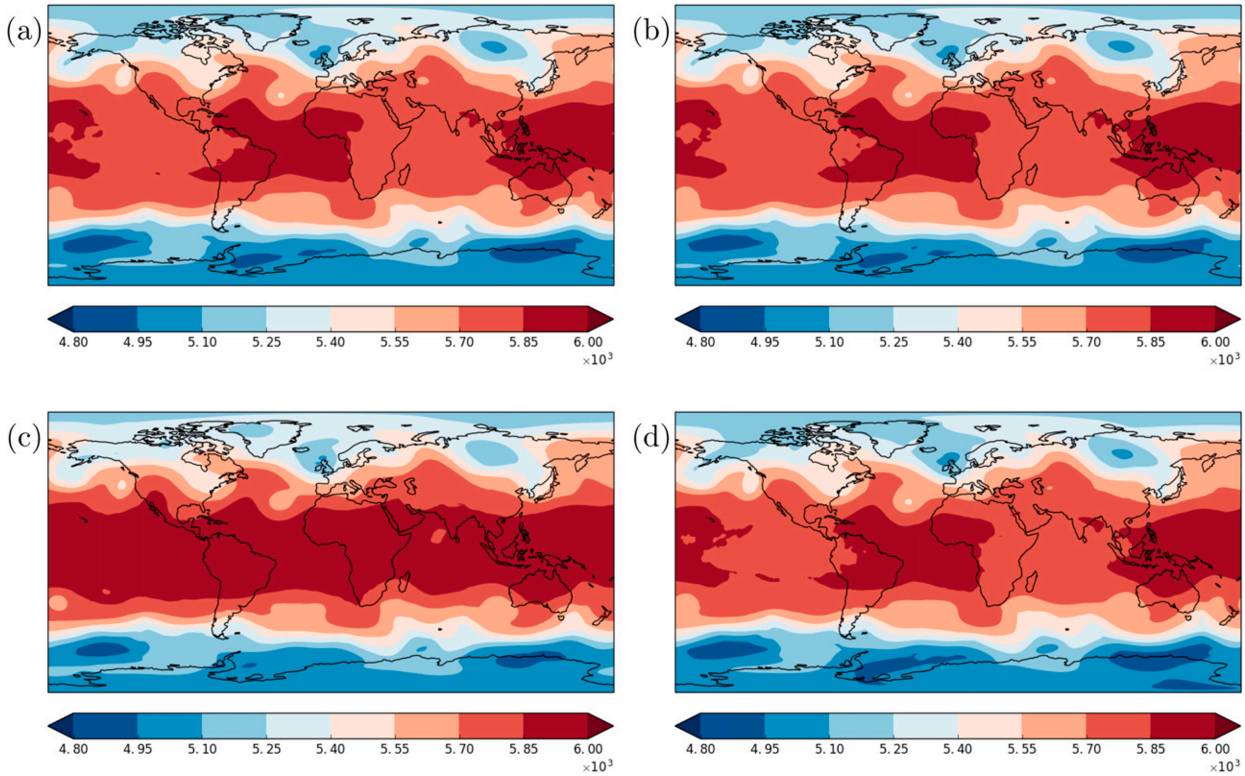

FIG. 2. Geopotential height of pressure level 500 hPa after 120 simulated hours. Differences between (a) double and (b) single are small, whereas (c) at 8 significand bit, variations can be seen globally, dominated by a global bias. (d) By retaining double precision for the zeroth total wavenumber (which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

spectral space involving the zeroth spectral mode, and 8 significand bits for all other modes. This produces a global mean Z500 of 5649.54 m, to be compared with double, 5648.47 m and single, 5648.79 m.

b. *Error measures*

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

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

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

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 is relative to the distance between double f_d and single-precision f_s integrations:

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

where nlev are the model levels. We consider Er for prognostic variables surface pressure, temperature, vorticity, and divergence at day 2. 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 grid points 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 grid points lie more than one standard deviation of the operational forecast ensemble from the double-precision forecast. This is measured over days 2–5 of

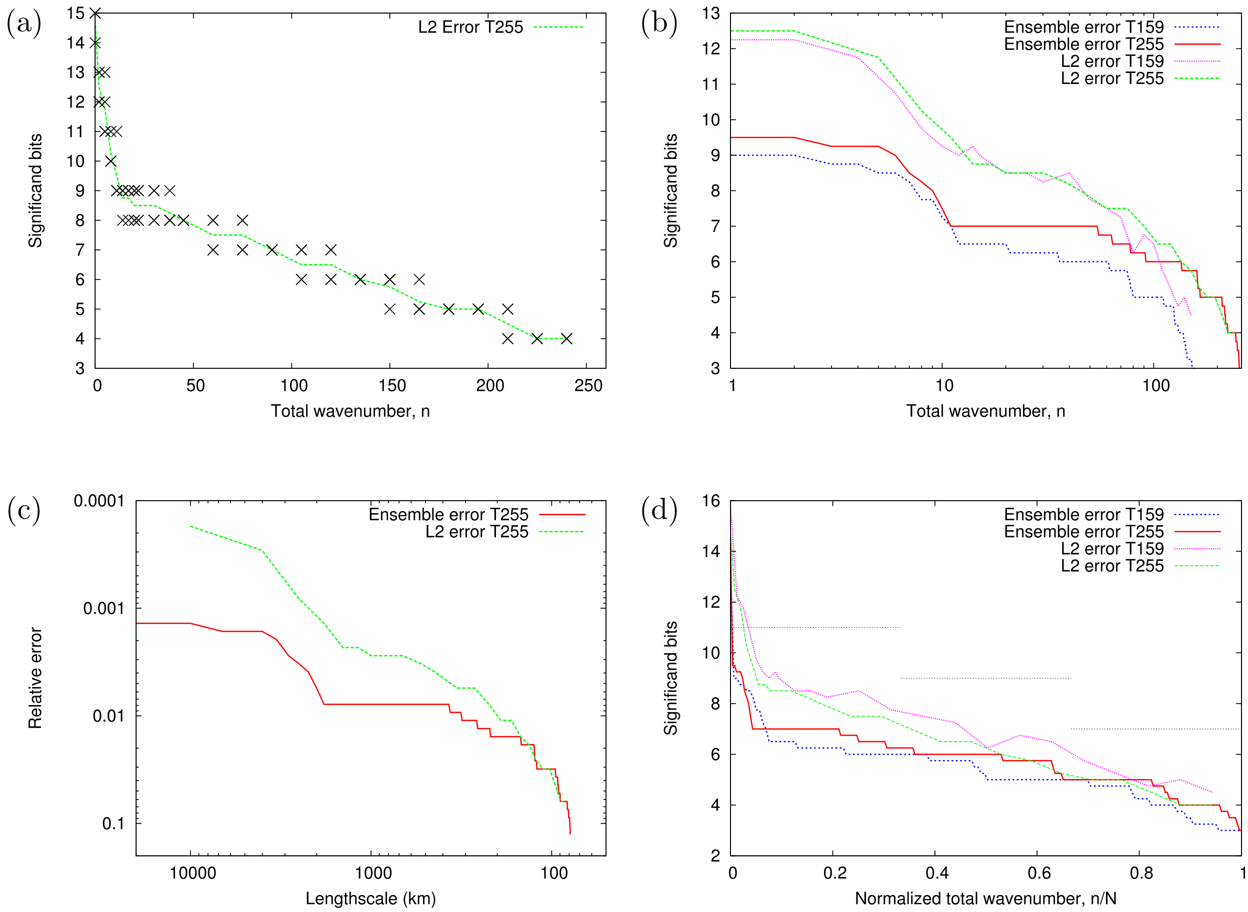

FIG. 3. (a) Precision needed to represent total wavenumbers $\geq n$ for an L_2 norm error less than two. Crosses represent four start dates at T255 resolution, with the line denoting the average. (b) Average precision for European ensemble error or L_2 error measures at resolutions T159 and T255. (c) Replotted T255 data as relative error against length scale. Please note that the x and y axes are reversed. (d) Total wavenumber is normalized by the truncation wavenumber to collapse the data for each measure. Gray dotted lines indicate scale-selective precision used for the decadal runs (see section 4d).

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 uncertainty and stochastic physics (ECMWF 2017), 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: 0000 UTC 26 October 2009, 0000 UTC 11 February 2010, 0000 UTC 26 October 2013, and 0000 UTC 11 February 2014. These dates cover two recent U.K. storm conditions in 2013 and 2014, as well as the same calendar date four years earlier that exhibited calm conditions for the United Kingdom.

c. *Optimal precision*

In Fig. 3a 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 toward lower precision being required at higher wavenumbers. A similar pattern is found when reducing precision only for a single total wavenumber (not plotted). Figure 3b plots the average precision against the wavenumber for both norms at two resolutions: T159 and T255.

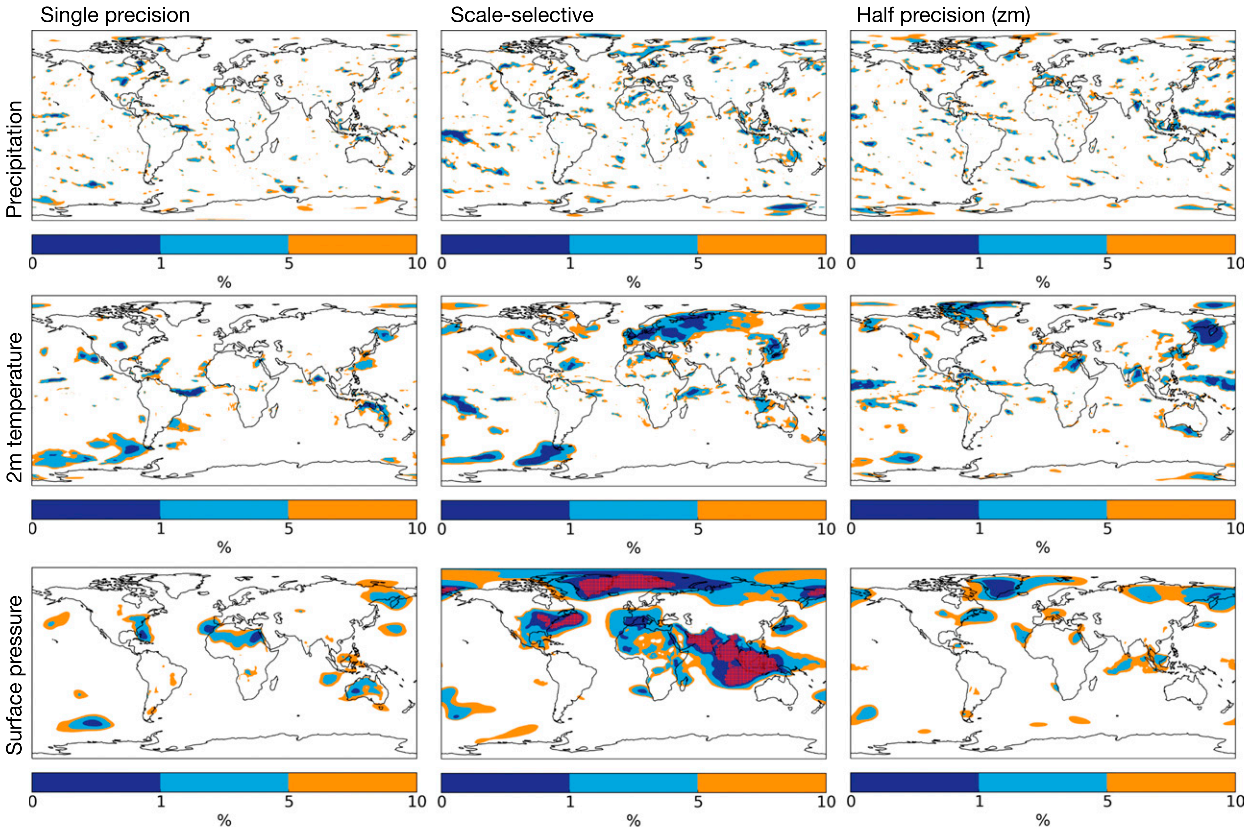

FIG. 4. Significance testing of decadal reduced-precision ensembles testing for differences to a double-precision ensemble for the accumulated precipitation, mean 2-m temperature, and mean surface pressure fields. Plotted are the local significance probabilities for single precision, weather-optimal scale-selective precision, and half precision (zeroth mode kept at double precision). Red dots indicate failed significance tests at the 5% level as calculated in Fig. 5.

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 Fig. 3c where the T255 data are transformed to length scale and relative error using the following formulas:

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

After rescaling the data by N, the truncation limit, we see good agreement between the two different resolutions (Fig. 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 require 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.

d. *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 1 to 10 January 2005 are integrated forward to the end of 2015 with prescribed observed SSTs. Discarding the data from 2005 we have a 10-member ensemble for the decade 2006–15. 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 based upon the results from weather prediction. Specifically, scale selective means double

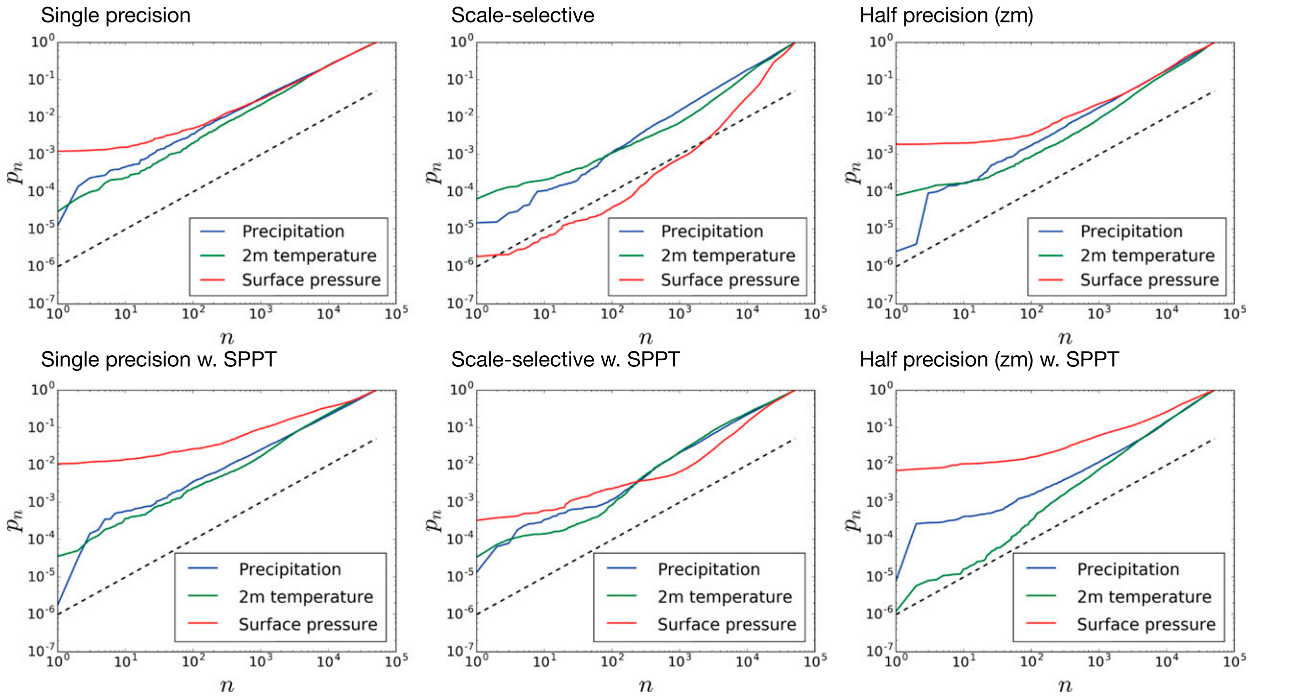

FIG. 5. (top) The local significance values from Fig. 4, ordered by size, to enable a global assessment of significance based upon the distribution of p values. The false discovery rate (FDR) test assesses a field as significantly different when values lie below the dashed-black curve (here denoting a test at the 5% level). All local tests with p values smaller than the largest failed test are considered significant and their locations are shown with red dots in the spatial significance maps of Fig. 4. For single precision and half precision, no fields are significantly different from double precision. The weather-optimal precision has a significantly different mean surface pressure. (bottom) The same test for ensembles with the stochastically perturbed parameterization tendency (SPPT) scheme active. With SPPT, no fields are significant at any of the tested precisions. Here scale-selective precision lies the furthest from being significant.

precision for the zero mode, 11 significand bits for total wavenumbers 1–50, 9 significand bits for the next 50, and 7 significand bits for the higher modes (see dashed gray lines in Fig. 3c). This is generally a cautious approach compared to our results, with the chosen precision being larger than that required for weather predictions with the exception of the first six (nonzero) modes. The importance of precision to these modes was found to decay with lead time, with 11 bits being acceptable beyond day 5. Half precision uses the 10-bit significand for all nonzero 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, 2-m temperature, and surface pressure from the double-precision 10-member ensemble. In Fig. 4, we plot maps of the probability implied by the t test at the 1%, 5%, and 10% levels. To test the significance of each field at each precision we use the false discovery rate (FDR) test (Benjamini and Hochberg 1995; Wilks 2006). This test examines the distribution of grid point p values, ordered by their magnitude, to find those that lie below the line described by $\mathrm{FDR}(n) = (n/N)p$, where n is the ordering index, N is the total number of tests (here the number of grid points), and p is the global p value for the test (here 5%). The distribution of p values for each field is shown in the top row of Fig. 5. For single precision and half precision (zm) all points lie above the FDR line and so no fields are found to be significantly different. For the scale-selective precision the mean surface pressure is found to significantly differ from double precision. All grid points with p values less than or equal to the largest failed p value are classified as significantly different and are denoted by red dots in Fig. 4.

The bottom row of Fig. 5 shows the FDR test applied to ensembles at varying precision with the stochastically perturbed parameterization tendency (SPPT) scheme active. This scheme, used in the ECMWF ensemble, applies multiplicative noise, with prescribed correlations in space and time, to the accumulated tendencies of the parameterization schemes (Buizza et al. 1999). The standard operational values for correlation length scales and time scales are used here (Shutts et al. 2011). This scheme has been shown to have a positive effect on both weather and climate time scales with not only increased

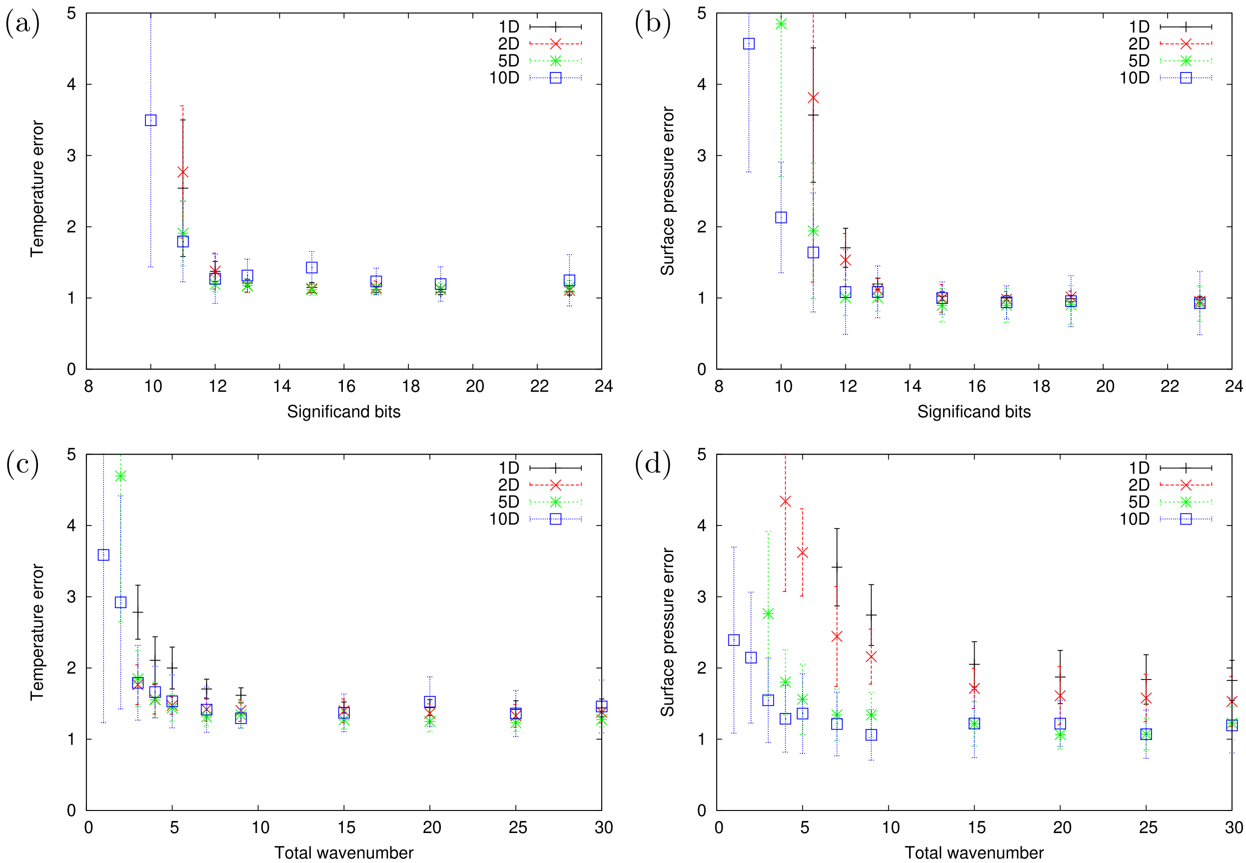

FIG. 6. Reduced-precision T511 experiments compared to the distance between double and single precision, averaged over 9 start dates after 1, 2, 5, and 10 days of integration. (a) Temperature error for uniform precision reduction in spectral space. (b) Surface pressure error for uniform precision reduction. (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.

ensemble spread but also changes in the mean state (Palmer et al. 2009; Weisheimer et al. 2014; Sanchez et al. 2016; Christensen et al. 2017). Examining each precision setup against a double-precision ensemble (also with SPPT active) we find that none of the measured fields are significantly different.

Establishing the optimal precision for climatological time scales will require more research. The results found here suggest that precisions far below single precision are feasible for integrations on long time scales. Measured against the spread of ensembles with SPPT, the differences induced by precision reductions are small.

e. *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 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 time step of 15 min. These start dates cover a range of months and conditions. The start dates considered are 0000 UTC 8 April 2011, 0000 UTC 27 October 2012, 0000 UTC 1 February 2013, 0000 UTC 15 March 2014, 1200 UTC 2 October 2015, 0000 UTC 7 March 2016, 0000 UTC 30 June 2016, 0000 UTC 25 August 2017, and 1200 UTC 5 September 2017. The effective grid resolution of approximately 39 km is closer to ECMWF's 20-km resolution (recently upgraded from 31 km 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 (Figs. 6a,b). Double precision is used for the zeroth mode. Equivalent plots for vorticity and divergence are comparable and are not shown here. As significand bits are decreased, the error remains comparable to single precision (close to 1) until 12 significand bits, below this the

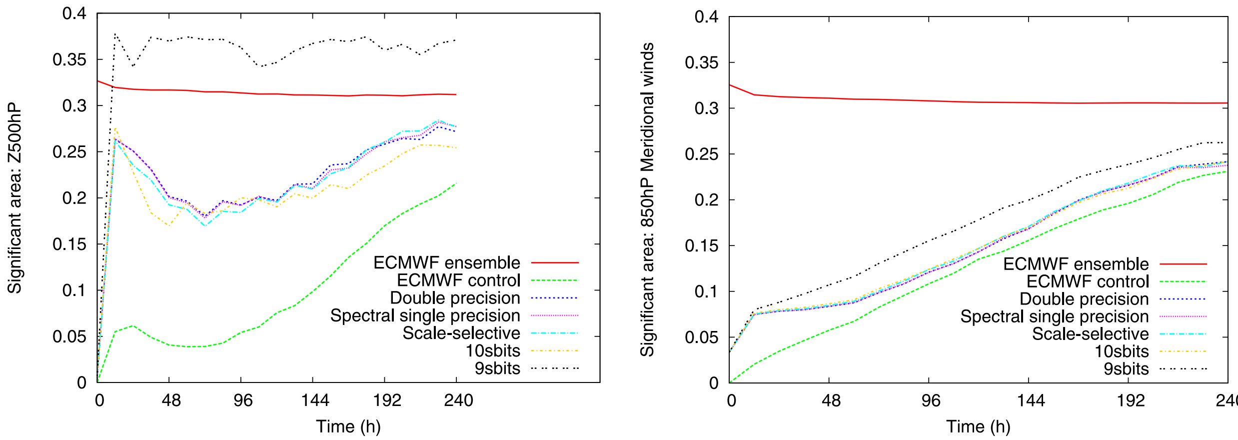

FIG. 7. Proportion of the globe that lies more than one ECMWF ensemble standard deviation from the ECMWF ensemble mean for (left) Z500 and (right) 850-hPa meridional winds. Members of the ECMWF ensemble remain close to a third, while the unperturbed ECMWF control climbs from zero. Our double-precision, 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.

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 Figs. 6c and 6d we plot the temperature and surface pressure errors for simulations that 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 can this be decreased to the fifth wavenumber. A model with 10 significand bits from wavenumber 5 onward 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. In Fig. 7 this measure is plotted against time for Z500 and 850-hPa 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 for 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. Conclusions

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 dependence, with large spatial scales requiring higher precision than small scales. Even at large scales the necessary precision is far below that of double precision.

Toward operational resolutions we continue to find that double precision is unnecessary, particularly when examined in the context of the operational ensemble. 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.

For decadal runs, the necessary precision again appears to be much lower than the double precision widely used. Computational constraints have prevented a full search for optimal precision over this time scale. However, both of the reduced-precision setups considered here are not significantly different from double precision when SPPT is used for the ensembles. The resolution dependence of precision on weather time scales appears to be approximately linear in truncation level, with low precision simply delayed to higher wavenumbers as resolution is increased. The success of half precision (with the zonal mode kept at double precision) on decadal time scales implies that resolution increases will have a limited impact on the necessary precision for these time scales.

While the reduction of precision is still limited to a fairly small fraction of the computations of the full model, we will continue to investigate a precision reduction in more model components. This work can 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 because of the matrix multiplications involved. Our work here suggests that half precision is plausible for these calculations. Investigating the spectral transforms will be carried out in future work. The inherent uncertainty in physical parameterization schemes makes these schemes a natural place for further reduced-precision studies.

Acknowledgments. M. Chantry was supported by a grant from the Office of Naval Research Global. P. D. Düben gratefully acknowledges funding from the Royal Society for his University Research Fellowship as well as funding from the ESIWACE project. ESIWACE has received funding from the European Union's Horizon 2020 research and innovation program under Grant 675191. T. N. Palmer received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Grant 741112). Acknowledgement is made for the use of ECMWF's computing and archive facilities in this research.

REFERENCES

Benjamini, Y., and Y. Hochberg, 1995: Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B*, **57**, 289–300.

Buizza, R., M. Milleer, and T. Palmer, 1999: Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **125**, 2887–2908, https://doi.org/10.1002/qj.49712556006.

Christensen, H., J. Berner, D. R. Coleman, and T. Palmer, 2017: Stochastic parameterization and El Niño–Southern Oscillation. *J. Climate*, **30**, 17–38, https://doi.org/10.1175/JCLI-D-16-0122.1.

Dawson, A., and P. D. Düben, 2017: rpe v5: An emulator for reduced floating-point precision in large numerical simulations. *Geosci. Model Dev.*, **10**, 2221–2230, https://doi.org/10.5194/gmd-10-2221-2017.

Düben, P. D., and T. N. Palmer, 2014: Benchmark tests for numerical weather forecasts on inexact hardware. *Mon. Wea. Rev.*, **142**, 3809–3829, https://doi.org/10.1175/MWR-D-14-00110.1.

ECMWF, 2017: IFS documentation Cy43r3. Part V: Ensemble prediction system. Tech. Rep., ECMWF, Shinfield Park, United Kingdom, 23 pp., https://www.ecmwf.int/sites/default/files/elibrary/2017/17737-part-v-ensemble-prediction-system.pdf.

Markidis, S., S. W. Der Chien, E. Laure, I. B. Peng, and J. S. Vetter, 2018: NVIDIA Tensor Core programmability, performance and precision. Proc. 2018 IEEE Int. Parallel and Distributed Processing Symp. Workshops, Vancouver, British Columbia, Canada, IEEE, 522–531, https://doi.org/10.1109/IPDPSW.2018.00091.

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

Palmer, T., R. Buizza, F. Doblas-Reyes, T. Jung, M. Leutbecher, G. Shutts, M. Steinheimer, and A. Weisheimer, 2009: Stochastic parametrization and model uncertainty. ECMWF Tech. Memo. 598, ECMWF, Shinfield Park, United Kingdom, 42 pp., https://doi.org/10.21957/ps8gbwbdv.

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

Sanchez, C., K. D. Williams, and M. Collins, 2016: Improved stochastic physics schemes for global weather and climate models. *Quart. J. Roy. Meteor. Soc.*, **142**, 147–159, https://doi.org/10.1002/qj.2640.

Shutts, G., M. Leutbecher, A. Weisheimer, T. Stockdale, L. Isaksen, and M. Bonavita, 2011: Representing model uncertainty: Stochastic parametrizations at ECMWF. ECMWF Newsletter, No. 129, ECMWF, Reading, United Kingdom, 19–24, https://doi.org/10.21957/fbqmkhv7.

Thornes, T., P. D. Düben, and T. N. Palmer, 2017: On the use of scale-dependent precision in earth system modelling. *Quart. J. Roy. Meteor. Soc.*, **143**, 897–908, https://doi.org/10.1002/qj.2974.

——, ——, and ——, 2018: A power law for reduced precision at small spatial scales: Experiments with an SQG model. *Quart. J. Roy. Meteor. Soc.*, **144**, 1179–1188, https://doi.org/10.1002/qj.3303.

Váňa, F., P. D. Düben, S. Lang, T. N. Palmer, M. Leutbecher, D. Salmond, and G. Carver, 2017: Single precision in weather forecasting models: An evaluation with the IFS. *Mon. Wea. Rev.*, **145**, 495–502, https://doi.org/10.1175/MWR-D-16-0228.1.

Weisheimer, A., S. Corti, T. Palmer, and F. Vitart, 2014: Addressing model error through atmospheric stochastic physical parametrizations: Impact on the coupled ECMWF seasonal forecasting system. *Philos. Trans. Roy. Soc. London*, **372A**, 2018, https://doi.org/10.1098/rsta.2013.0290.

Wilks, D., 2006: On "field significance" and the false discovery rate. *J. Appl. Meteor. Climatol.*, **45**, 1181–1189, https://doi.org/10.1175/JAM2404.1.

Zuras, D., and Coauthors, 2008: IEEE standard for floating-point arithmetic. IEEE Std. 754–2008, 70 pp., https://doi.org/10.1109/IEEESTD.2008.4610935.