Accelerating high-resolution weather models with deep-learning hardware

Sam Hatfield
Atmospheric, Oceanic and Planetary Physics, University of Oxford
Oxford, UK
samuel.hatfield@physics.ox.ac.uk

Matthew Chantry
Atmospheric, Oceanic and Planetary Physics, University of Oxford
Oxford, UK
matthew.chantry@physics.ox.ac.uk

Peter Düben
European Centre for Medium-Range Weather Forecasts
Reading, UK
peter.dueben@ecmwf.int

Tim Palmer
Atmospheric, Oceanic and Planetary Physics, University of Oxford
Oxford, UK
tim.palmer@physics.ox.ac.uk

ABSTRACT

The next generation of weather and climate models will have an unprecedented level of resolution and model complexity, and running these models efficiently will require taking advantage of future supercomputers and heterogeneous hardware.

In this paper, we investigate the use of mixed-precision hardware that supports floating-point operations at double-, single- and half-precision. In particular, we investigate the potential use of the NVIDIA Tensor Core, a mixed-precision matrix-matrix multiplier mainly developed for use in deep learning, to accelerate the calculation of the Legendre transforms in the Integrated Forecasting System (IFS), one of the leading global weather forecast models. In the IFS, the Legendre transform is one of the most expensive model components and dominates the computational cost for simulations at a very high resolution.

We investigate the impact of mixed-precision arithmetic in IFS simulations of operational complexity through software emulation. Through a targeted but minimal use of double-precision arithmetic we are able to use either half-precision arithmetic or mixed half/single-precision arithmetic for almost all of the calculations in the Legendre transform without affecting forecast skill.

CCS CONCEPTS

• **Applied computing → Earth and atmospheric sciences**.

KEYWORDS

Numerical weather prediction, spectral models, Legendre transforms, floating-point arithmetic, half-precision, Tensor Core

PASC '19, June 12–14, 2019, Zurich, Switzerland

ACM ISBN 978-1-4503-6770-7/19/06…$15.00
https://doi.org/10.1145/3324989.3325711

ACM Reference Format:
Sam Hatfield, Matthew Chantry, Peter Düben, and Tim Palmer. 2019. Accelerating high-resolution weather models with deep-learning hardware. In *Proceedings of the Platform for Advanced Scientific Computing Conference (PASC '19), June 12–14, 2019, Zurich, Switzerland.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3324989.3325711

1 INTRODUCTION

High-performance computing architectures are currently undergoing a paradigm shift, prompting a reconsideration of how we design scientific computing applications [9]. Future supercomputers will be heterogeneous, employing a variety of computing hardware, and will place more emphasis on data transfer between individual computing units, including memory-to-processor and node-to-node communication, compared with the traditional focus on raw floating-point computation. This presents a challenge to architects of numerical weather prediction (NWP) models and threatens the continual climb of weather forecast skill, maintained over the past few decades, which provides significant economic benefits to society [1].

Of the alternatives to traditional central processing units (CPUs), graphics processing units (GPUs) are likely to see an increasing presence in supercomputing centres. These devices allow a far greater degree of parallelism over a multi-core CPU and have seen a surge in popularity within science over the past few years. Notable achievements within NWP include Fuhrer et al. [5] and Yashiro et al. [19], who both successfully ported existing non-hydrostatic models to run on several thousand GPUs. The latest GPUs come with several features aimed at accelerating common computations within deep-learning applications, notably matrix-matrix multiplications. The NVIDIA V100, for example, features a separate matrix-matrix multiplier known as a "Tensor Core" for which they claim a speed-up factor of 16 times compared with a standard double-precision matrix-matrix multiply [13]. However, this device can only operate on half-precision floating-point numbers (the output can be half-precision or single-precision) so any potential application must be tolerant to the relatively large rounding-errors incurred by putting the input matrices into half-precision containers. In this paper, we consider an application for these reduced-precision matrix-matrix multipliers within numerical weather prediction models.

A number of other studies have investigated the use of floating-point arithmetic below single-precision in weather and climate models. For simple models it is feasible to run the entire model at a significantly reduced precision. Thornes et al. [16] and Hatfield et al. [7], for example, both used half-precision for all or almost all calculations. For intermediate or high complexity models, however, it is usually necessary to secure certain portions of the model at the standard double-precision to avoid deleterious errors, such as overflows and catastrophic cancelling errors. Nevertheless, Düben et al. [4] were able to perform 84% of the floating-point calculations in a low resolution atmospheric dynamical core with only 6 bits for the floating-point significand. While the use of single-precision has recently been tested successfully in several weather and climate models, the uses for half-precision floating-point numbers are limited because of the small exponent width and therefore small range of representable numbers.

There are different methods to discretise the equations of motion of the atmosphere for weather and climate models. The "spectral transform technique" has been the method of choice at the European Centre for Medium-Range Weather Forecasts (ECMWF) almost since its inception [15]. In a spectral model, a three-dimensional scalar meteorological field, such as temperature, is discretised vertically into a series of spherical shells extending from the Earth's surface to the upper atmosphere. Each spherical shell is then formulated as a weighted sum of spherical harmonic basis functions. The applied discretisation is similar to a Fourier decomposition on the plane and we refer to the space spanned by these basis functions as "spectral space". The use of a spectral model opens opportunities for the use of mixed-precision. Chantry et al. [2] developed a "scale-selective" approach to reducing-precision, in the spectral space calculations of the Integrated Forecasting System (IFS) of ECMWF: they used a high precision for large-scale atmospheric variables and a low precision for small-scale atmospheric variables. Overall this allowed them to use far fewer bits than the baseline, global double-precision, without affecting the weather forecasting skill.

Time-stepping and computation of derivatives, among other things, are efficiently computed in spectral space. Computation of the Laplacian of a field, for example, is trivial in spectral space. However, some quantities, such as tendencies due to parametrised physical processes, must still be computed on the grid points of a common grid in physical space, referred to as "grid point space". This requires a transform of several fields from spectral space to grid point space and back every timestep. Each transform requires the use of a fast Fourier transform and a Legendre transform. In the past, the transforms did not constitute a significant fraction of the total computational cost of the model and so this seemingly unwieldy process paid off. However, the Legendre transform scales poorly and the cost of this process tends to dominate the total cost of the system as horizontal resolution is increased [17]. At the highest operational resolution (9 km), the transforms back and forth between spectral and grid point space already constitute around 40% of the total computational cost (atmosphere only and without I/O). This problem has motivated the development of more efficient Legendre transform algorithms than the standard matrix-matrix multiplication, which scales as $O(n^3)$, where n is proportional to the number of latitudes and longitudes and therefore denotes the model's horizontal resolution. The "fast Legendre transform", for example, scales as $O(n^2 \log n)$ [18]. However there is a large minimum cost to this algorithm such that it only becomes competitive at very high resolutions and therefore the fast Legendre transform is still not used in operational forecasts at ECMWF.

In this paper, we explore the potential of GPUs to accelerate the traditional Legendre transform. In particular, we assess the impact of using half-precision matrix-matrix multiplications and the aforementioned Tensor Core on the meteorological skill of the ECMWF model, the IFS. With a few simple treatments we find that both half-precision arithmetic and the Tensor Core can be used successfully. We focus on the Legendre transform for the aforementioned reasons but also because it has been identified as one of the key "dwarfs" that should be targeted for optimisation when porting existing weather models to future heterogeneous architectures (specifically, the Fourier transform *and* the Legendre transform constitute the dwarf) [12]. However, the techniques outlined in this paper could in principle be used to accelerate any matrix multiplication, which is a standard operation used in many different components of weather forecast models.

In Section 2 we introduce the Legendre transforms mathematically and algorithmically, in Section 3 we discuss the implications for reducing precision in the Legendre transforms, in Section 4 we evaluate the impact of this procedure on the skill of weather forecasts, both deterministic and probabilistic, in Section 5 we estimate the computational cost of reduced-precision Legendre transforms and finally we conclude in Section 6.

2 THE LEGENDRE TRANSFORM

Any real horizontal scalar field on the sphere $f^p(\lambda, \theta)$, where λ is longitude, θ is latitude and p indexes the field (i.e. which model level and meteorological variable), can be written as a sum of spherical harmonics, $Y_{m,n}(\lambda, \theta)$:

$$f^p(\lambda, \theta) = \sum_{m=-\infty}^{\infty} \sum_{n=|m|}^{\infty} \psi^p_{m,n} Y_{m,n}(\lambda, \theta) \tag{1}$$

where $\psi^p_{m,n}$ is the complex amplitude of the (m, n) spectral mode for field p, m is the zonal (or "Fourier") wavenumber and n is the total wavenumber. Spherical harmonics are the eigenfunctions of the Laplacian in spherical coordinates and can be written as a product of a function of latitude and a function of longitude:

$$Y_{m,n}(\lambda, \theta) = P_{m,n}(\mu) e^{im\lambda}, \tag{2}$$

where $P_{m,n}(\mu)$ is an associated Legendre polynomial of the first kind and $\mu = \cos(\theta)$. In practice, the sums in Equation 1 are truncated at some total wavenumber N and zonal wavenumber M, which are the same in the case of the typical "triangular truncation". The truncation wavenumber N determines the model resolution which is referred to by the TXN labelling convention, "T" meaning "triangular" and "X" referring to the type of grid used in grid point space. We consider both the linear reduced Gaussian grid (TLN) and the cubic octahedral reduced Gaussian grid (TCoN), the latter now being used operationally at ECMWF. A spectral resolution of TCo639 corresponds to a grid point resolution of approximately 18 km at the Equator, whereas TCo3999 corresponds to a resolution of approximately 2.5 km. To transform from grid point space to

spectral space, a one-dimensional Fourier transform is applied to generate a Fourier representation of the fields along each longitude. In a second step, a Legendre transform is applied to obtain the spectral coefficients $\psi^p_{m,n}$ for the global representation in spectral space. To transform from spectral space to grid point space, an inverse Legendre transform is applied followed by an inverse Fourier transform.

Applying the Fourier transform to a grid point field produces the Fourier amplitudes at each latitude, $F^p_m(\mu)$. The direct Legendre transform then completes the transform into spectral space:

$$\psi^p_{m,n} = \int_{-1}^{1} F^p_m(\mu) P_{m,n}(\mu) d\mu. \tag{3}$$

Equation 3 is evaluated numerically using Gaussian quadrature:

$$\int_{-1}^{1} F^p_m(\mu) P_{m,n}(\mu) d\mu = \sum_{i=1}^{N+1} w_i F^p_m(\mu_i) P_{m,n}(\mu_i), \tag{4}$$

where the index i iterates over the Gaussian $N+1$ latitudes, w_i are the Gaussian weights and μ_i is the location of the ith Gaussian latitude. The inverse Legendre transform is defined by

$$F^p_m(\mu) = \sum_{n=|m|}^{N} \psi^p_{m,n} P_{m,n}(\mu). \tag{5}$$

In the IFS, equations 4 and 5 are computed as a series of matrix-matrix multiplications parallelised over each m. In practice, the exact operations differ from those presented so far due to several complications. Firstly, the spectral amplitudes are complex and so the above transforms must be performed separately for the real and imaginary components. Secondly, the cost of both the direct and inverse transforms can be reduced by a factor of two by taking advantage of the symmetry properties of the associated Legendre polynomials. Finally, because the meteorological variables are all real, only the modes with a positive zonal wavenumber m must be stored (the negative m spectral coefficients must be the exact complex conjugate of the corresponding positive m coefficients).

3 REDUCING PRECISION IN THE LEGENDRE TRANSFORMS

When reducing the precision of a floating-point operation, two effects can be anticipated. The first is an increase in the rounding error that occurs. According to the "standard model" of floating-point arithmetic [8], the result of the floating-point equivalent of an arithmetical operation is the exact result with a multiplicative error, $(1+\delta)$, where δ is the rounding error bounded by $|\delta| \leq \epsilon$ and ϵ is the machine epsilon. For example, the floating point equivalent of $x + y$ is

$$\mathrm{fl}(x+y) = (x+y)(1+\delta). \tag{6}$$

The machine epsilon is defined by $\epsilon = 2^{-p-1}$ where p is the number of significand (also known as the mantissa) bits. For each significand bit that is removed, ϵ doubles and therefore larger rounding errors become possible. The overall error of a particular model, consisting of many millions of floating-point operations, will therefore also increase. However, this error increase must always be considered with comparison to the margin of uncertainty provided by the inherent uncertainties in the modelling formulation. Increased rounding-errors from a precision-reduction will not necessarily be noticeable, in the presence of model error.

The second effect comes from a reduction in the exponent width. The exponent determines the range of representable numbers. For example, the maximum double-precision number (11 exponent bits) is around 10^{308} whereas the maximum half-precision number (5 exponent bits) is only 65504. If a number larger than this maximum is stored in a floating-point variable, or results from a floating-point operation, an overflow occurs and the number is rounded to infinity which typically results in the model crashing. Similarly, any numbers smaller than the smallest possible floating-point number will underflow and be rounded to zero. This will only crash the model if the resulting number is used as a divisor leading to a divide-by-zero error. Here, we focus only on the matrix-matrix multiplications used for the Legendre transforms, so we prioritise the avoidance of overflow errors instead of underflow errors, because no division operations are used.

Naively using half-precision arithmetic to perform the Legendre transforms will lead almost instantly to a crash of the model. This is because numbers larger than the maximum half-precision value occur, are rounded to infinity and then infect every calculation that they occur in eventually causing a floating-point exception. However, a simple procedure can secure the Legendre transforms from overflow errors. Given that equations 4 and 5 are linear operations, by multiplying the field-to-be-transformed by a scalar and dividing the transformed-field by the same scalar we recover the correct result had we applied no rescaling. This technique was used by Micikevicius et al. [11] in order to allow mixed-precision training of a neural net. We propose to use this simple procedure to enable half-precision computations for the Legendre transforms.

For both the forward and inverse transforms, we first compute the maximum of the incoming field. We then multiply the entire field by a rescaling factor such that this maximum is equal to a prescribed value a. We then apply the transform using either a half-precision matrix-matrix multiplication (half_trans) or the Tensor Core (tensor_core) and divide the result by the rescaling factor.

We performed tuning experiments to determine the optimal value for a by computing the day 5 root-mean-square error of 500 hPa geopotential height with respect to a fully double-precision simulation for different values of a between 0.1 and 1000 (values outside this range gave significantly larger errors or crashed entirely, likely due to overflows and underflows). We found no change in the error within this range. This is not surprising given that floating-point numbers have a constant relative precision — within the range of representable numbers there is no "optimal rescaling" that will minimise the error. We decided to use $a = 100$ as this happened to give a slightly lower error than other values, but we do not consider this choice significant. Note that it is technically possible (but unlikely) for overflows to occur when $a = 100$ for high resolutions. This is because the longest sum for a resolution of spectral truncation N is $N+1$ elements long, and therefore if all elements of this sum happen to be close to a then the result of the sum will be larger than 65504.

To test the use of mixed-precision with IFS without porting the model code to GPUs we use a software emulator to emulate half-precision arithmetic, specifically the Reduced Precision Emulator [3]. As a consequence we are not able to directly measure

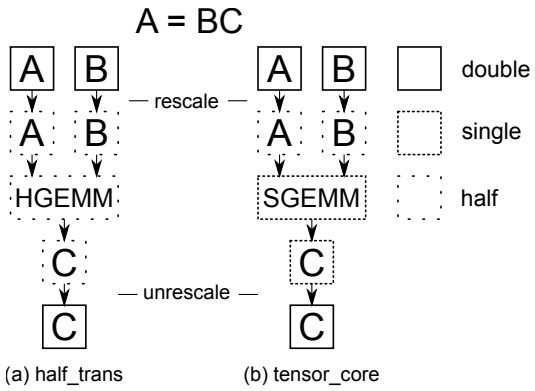

Figure 1: A visualisation of the two core experimental setups: (a) half_trans and (b) tensor_core.

the speed-up that would be obtained from running the code on the actual target hardware. Nevertheless, it is important to assess the impact of the proposed hardware changes on the forecast skill of the model, and software emulation allows us to do this on standard hardware with minimal modifications to the code.

We investigate the difference between the matrix-matrix multiplication in two reduced-precision setups, half_trans and tensor_core, that are illustrated in Figure 1. For both setups, the double-precision inputs A and B are first rescaled and then placed into half-precision containers. Then, the matrix-matrix multiplication is performed at half-precision for half_trans (HGEMM) and single-precision for tensor_core (SGEMM). This is a realistic emulation of the Tensor Core on an NVIDIA V100 GPU that takes half-precision inputs but produces a single-precision output. We verified this emulation procedure against a real Tensor Core on an NVIDIA V100 GPU. We computed a matrix-matrix multiplication both using the emulator and using an actual Tensor Core and found that the difference between the two matrices was consistent with standard matrix-matrix multiplication error bounds [8]. Finally, for both setups, the output is then "unrescaled" and placed into a double-precision output container.

When we applied the aforementioned scaling technique to the Legendre transforms, we encountered a further problem arising from the spectral modes with a zonal wavenumber (m) of zero. The spectral coefficient $\psi^p_{0,0}$ is by definition the global mean of the field p and so, for certain fields such as temperature, this has a much greater amplitude than the remaining coefficients. Therefore, when the $F_0(\mu)$ term is formed from the sum in Equation 5 (the inverse Legendre transform) rounding errors can lead to a loss of precision and spurious zonally-symmetric patterns in the transformed fields. A similar effect is observed for the direct Legendre transform, though the exact reason why is not clear to us. The direct transform consists of a sum over latitude and it is not obvious that one of the terms in this sum should dominate the others in magnitude in a similar way to the inverse transform. In any case, we can eliminate this problem simply by using double-precision arithmetic for the $m = 0$ direct and inverse transforms. The remaining zonal wavenumbers are still computed using half-precision and so this fix has a negligible impact on the total cost of the transforms. Additionally, the model initialisation was still performed with double-precision arithmetic, even if the Legendre transforms used half-precision arithmetic. Several transforms are performed before the main integration begins which we reason should be performed with a high-precision.

4 FORECAST TESTS

At ECMWF, two different kinds of weather forecasts are generated. For "deterministic forecasts" a single model simulation is started at the highest resolution possible to generate a forecast that provides the most likely state of the weather in the future. However, it is often very important to also provide probability distributions for predictions. Therefore, a "probabilistic forecast" is performed that uses an ensemble of fifty simulations that are slightly perturbed so that the variability of the different predictions can be used as a measure of forecast uncertainty (a large spread between ensemble members indicates that a prediction is not reliable). In the following, we will study both kinds of predictions.

A reduction of numerical precision will inevitably reduce forecast quality. However, weather and climate forecasts are perturbed by several sources of errors (imperfect initial conditions, errors in the formulation of the model, limited resolution that is insufficient to represent all important processes of the Earth System etc.). It is therefore possible that errors due to reducing precision will be insignificant in comparison to other errors. To test whether rounding errors are significant in our mixed-precision simulations, we compare the results with simulations that use a stochastic parametrisation scheme. These schemes are designed to generate spread between different simulations of an ensemble forecast by adding a stochastic forcing into the model configuration. Since this spread was adjusted to fit the average error of weather forecasts, the difference between model simulations with and without the stochastic parametrisation scheme enabled can serve as an estimate of the magnitude of model error. For our tests we used the stochastically perturbed parametrisation tendencies (SPPT) scheme which applies a multiplicative stochastic forcing to the total tendencies provided by the physics parametrisations [14]. As long as the impact from reducing precision is smaller than the impact of SPPT, reducing precision is likely to not noticeably affect the probabilistic skill of the model.

4.1 Deterministic forecast skill

As a first test we performed a double-precision 10 day simulation and compared it with several setups. For the experiments in this section we used the open-source variant of the IFS, OpenIFS, at cycle 38r1. This cycle was introduced in 2012 and used a linear grid. For each setup we computed the global mean root-mean-square error of the 500 hPa geopotential height field (Z500 RMSE) of the setup with respect to the double-precision control. We used the exact same initial conditions for all forecasts in this section so that only model error with respect to the double-precision control would cause two forecasts to diverge. For the first setup we used half-precision Legendre transforms (half_trans) and for the second setup we used the emulated Tensor Core (tensor_core), as discussed in Section 3. Our third setup consisted of a model with entirely

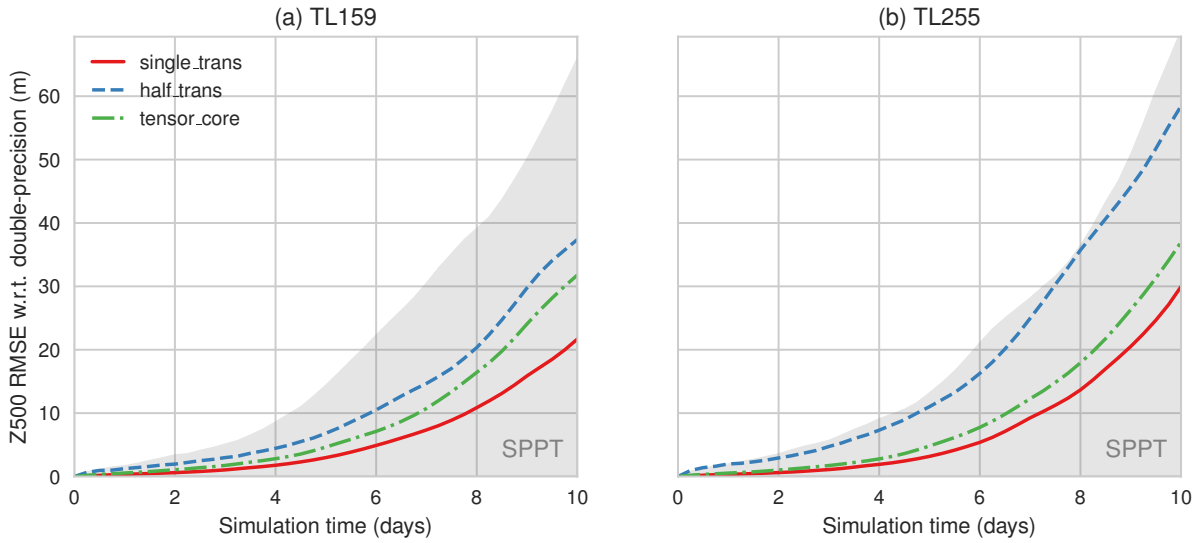

Figure 2: The root-mean-square error of the 500 hPa geopotential height field for various experimental setups with respect to a double-precision control experiment at (a) TL159 resolution and (b) TL255 resolution. The grey shaded area shows the error from a double-precision simulation with SPPT.

single-precision Legendre transforms (single_trans). Our final setup consisted of an entirely double-precision model with SPPT.

The results of this experiment for resolutions of TL159 and TL255 are shown in Figure 2. In this case we performed simulations for four different start dates and averaged the results. The lowest error occurred for single_trans, followed by tensor_core, half_trans and finally SPPT. The latter is displayed as a grey shaded area. For both resolutions, the error introduced by using half-precision or Tensor Core Legendre transforms was lower than that introduced by using stochastic physics.

We also performed experiments with the same setups but at TL511 resolution and for only one date. However, we found that the error when using half-precision Legendre transforms was substantially higher even than the SPPT setup. To remedy this, it was necessary to keep the computation of several more zonal wavenumbers at double-precision, instead of just the first as in the previous half-precision experiments. Figure 3 shows the effect of computing the first c wavenumbers using double-precision, with c ranging from 1 (as in the previous experiments) up to 20. Here we only simulated one hour. Evidently, c is a parameter that can be tuned to reduce the error to any desired level. However, the larger the value of c, the fewer operations are performed at half-precision and therefore the lower the cost saving. We chose to keep $c = 10$ more zonal wavenumbers at double-precision as the error for this setup is of the same order of magnitude as the error of SPPT (at least over the first hour, over which the impact of SPPT is known to be very small). When $c = 10$, only around 5% of the total number of floating-point operations are computed at double-precision. Keeping low wavenumbers, which represent large-scale atmospheric motions, at a high-precision is similar to the scale-selective approach advocated by Chantry et al. [2], though they did not consider the Legendre transform.

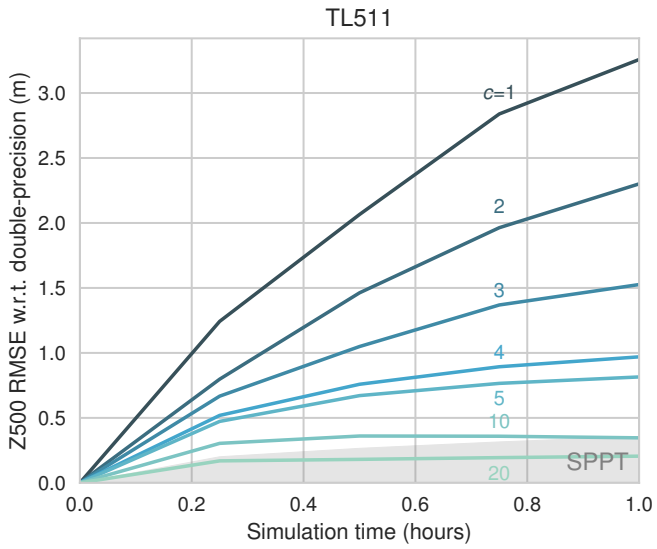

Figure 3: The root-mean-square error of the 500 hPa geopotential height field for different choices of the cutoff parameter c when using half-precision Legendre transforms and for a double-precision simulation with SPPT enabled.

Figure 4 shows results from the 10 day TL511 simulations. The effect of performing the first 10 zonal wavenumber computations at double-precision (half_trans_10) instead of just the first (half_trans) is immediately noticeable. Note that, as of July 2018, the resolution of the operational ensemble prediction system of ECMWF is TCo639 (our version of OpenIFS is not strictly comparable with the

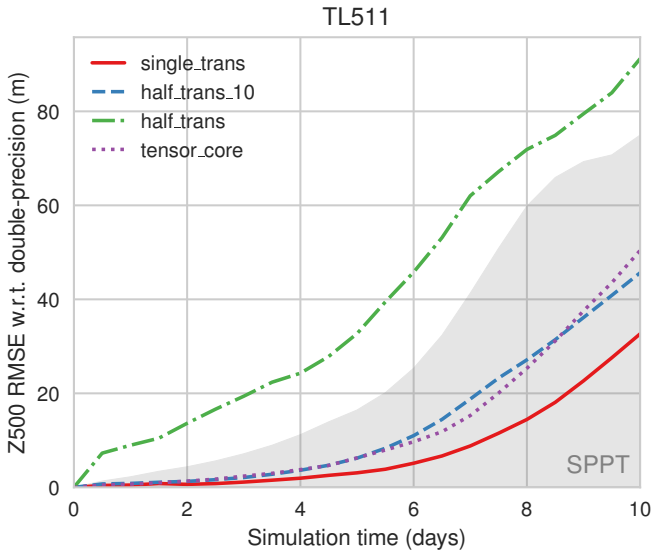

Figure 4: The root-mean-square error of the 500 hPa geopotential height field for various experimental setups with respect to a double-precision control experiment at TL511 resolution.

operational IFS as the latter uses a cubic grid whereas ours uses a linear grid). In the next section, we will evaluate the performance of the half_trans_10 and tensor_core setups in an operational-like context.

4.2 Probabilistic forecast skill

We performed a series of ensemble forecasts over 2017 with 12 start dates each separated by 30 days, beginning on January 1st. For the experiments in this section we used the IFS at cycle 43r3 and used a cubic octahedral grid. This cycle was used operationally at ECMWF from July 2017 until June 2018. We used 11 ensemble members for each forecast: 1 control member and 10 members with perturbed initial conditions and SPPT enabled. We performed the forecasts at TCo399 resolution which is roughly 50 km resolution at the Equator. To verify these forecasts, we used the so-called ECMWF "analysis" as a reference. The analysis is produced by combining observations with the IFS through data assimilation and therefore represents the best guess of the actual atmospheric conditions that occurred at the verification time. Unlike the observations, the analysis is defined on the same grid as the forecast and is therefore a convenient reference for computing error scores. We verified the forecasts using the continuous ranked probability score (CRPS) corrected for the limited ensemble size [10] and averaged this score over all forecasts. CRPS is a standard diagnostic to measure the quality of ensemble predictions. A low CRPS is desirable as it indicates an ensemble with a low spread and accurate mean. We considered three setups: double-precision (double), half-precision Legendre transforms for wavenumbers 10 up to 399 (half_trans_10) and Tensor Core Legendre transforms (tensor_core).

The results are illustrated in Figure 5 for three latitude bands, the northern and southern extratropics and the tropics, and for the 500 hPa geopotential height and temperature fields. For all latitude bands and both fields, the half-precision setup was competitive with the double-precision setup. The Tensor Core setup, on the other hand, demonstrated a slight average increase in the forecast error. This possibly indicates that the first 10 zonal wavenumbers are also important for tensor_core and should therefore be promoted to double-precision, as they are for half_trans_10.

As shown in Figures 2 and 4, reducing precision in the Legendre transforms has a clear impact on the model. However, this difference is hidden when considering the skill of probabilistic forecasts that are perturbed by both errors in the initial condition and model formulation. This demonstrates that, whenever assessing a model change, it is not sufficient to consider only deterministic forecast verification metrics in a perfect model context. This is also consistent with the findings of Hatfield et al. [6], who considered a precision reduction within a data assimilation model.

4.3 Operational resolution deterministic forecast skill

We also performed high-resolution deterministic forecasts for the same three setups as in Figure 5. Again, we used the IFS at cycle 43r3. We used 6 dates for the double and tensor_core setups but only 3 dates for the half_trans setup, due to computational budget limitations. We ran these experiments at the same resolution as the operational deterministic forecast at ECMWF at the time of writing, TCo1279. This is approximately 9 km resolution at the Equator. As in Section 4.2 we verified with respect to analysis and therefore, unlike in Section 4.1, these forecasts did include initial condition and model error. The deterministic forecast tests in this section are therefore a more realistic assessment of our proposed model change. For the half_trans setup we decided to secure the first 25 zonal wavenumbers with double-precision arithmetic, instead of 10 for the TL511 and TCo399 experiments. We reasoned that the optimal ratio of wavenumbers computed at double-precision to those computed at half-precision (2:100 for TL511) is not likely to change as resolution is increased. Accordingly, we computed $c = 25$ out of 1280 wavenumbers using double-precision with half-precision for the remainder so this setup is labelled half_trans_25. However, we did not have sufficient computational resources to tune this value so we do not claim that it is optimal.

Figure 6 shows the results from these forecasts for the 500 hPa geopotential height and temperature fields, averaged across all available dates for each setup. There were no problems with numerical stability for the tensor_core and half_trans_25 setups and the forecast skill is competitive with double-precision. In the extratropics there doesn't appear to be a significant difference between the three setups (the differences for half_trans_25 are probably due to using fewer start dates). However, the error for half_trans_25 and tensor_core seems to be slightly increased for Z500 in the tropics. This indicates that we should perform tests with tensor_core that also compute the leading zonal wavenumbers in double-precision.

As a final test we performed a forecast of Hurricane Irma which caused significant damage in the Caribbean and the southeastern United States in September 2017. For computational budget reasons

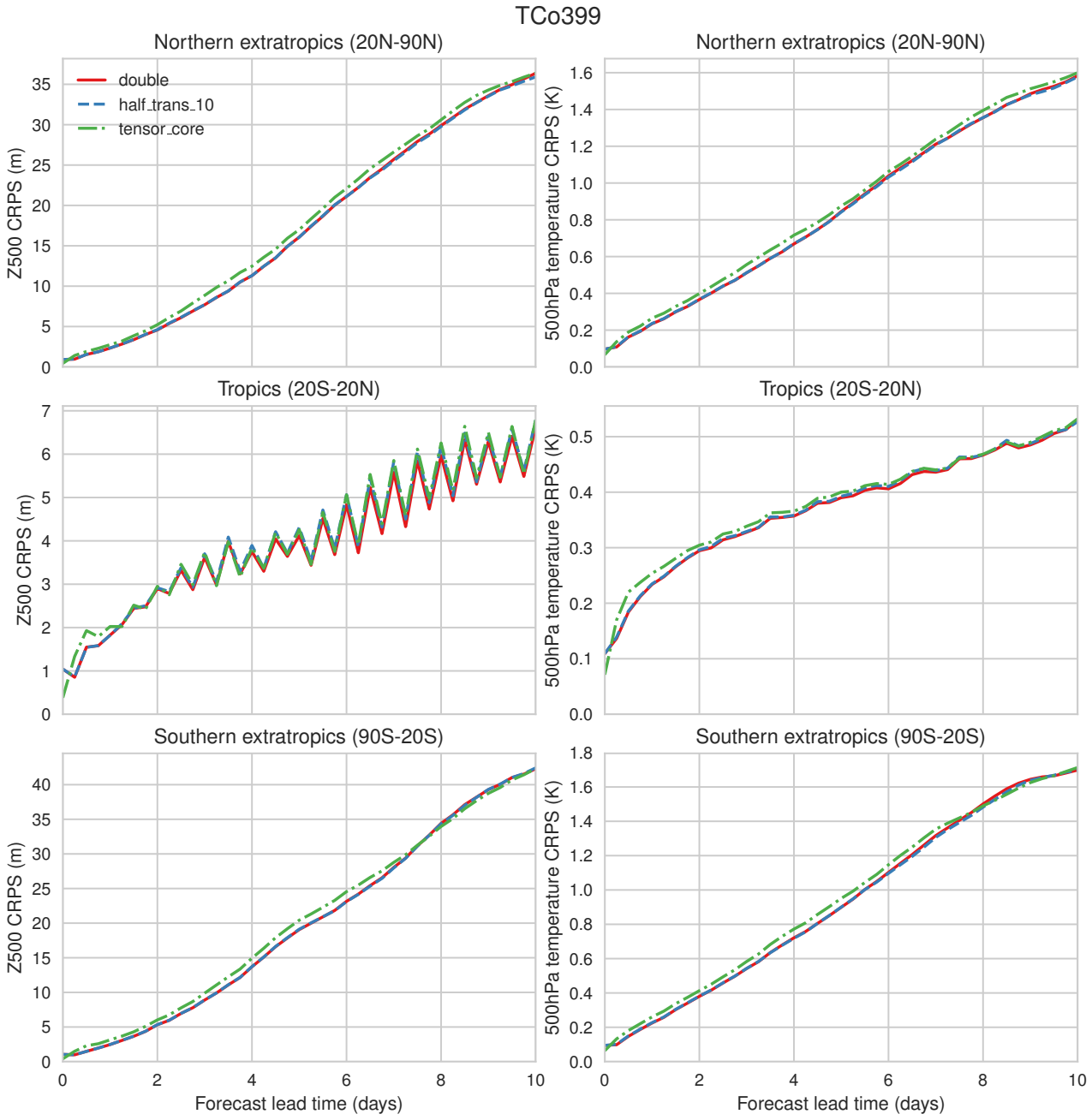

Figure 5: The continuous ranked probability score (CRPS) of the 500 hPa geopotential height and temperature fields for various experimental setups verified with respect to analysis at TCo399 resolution. The CRPS was averaged over 12 forecasts and there were 11 ensemble members for each forecast.

we only compared the double and tensor_core setups. We performed this forecast at TCo1279 resolution beginning on the 5th September 2017 at 12:00 pm. Note that the hurricane is already present in the initial conditions, having begun forming 5 days earlier. We simply wish to assess the ability of the tensor_core setup to match the trajectory forecasted by the state-of-the-art double-precision model.

The results from this experiment are given in Figure 7. The mean sea-level pressure at day 5 (12:00 pm on September 10th) is shown along with the simulated hurricane tracks for the analysis and the

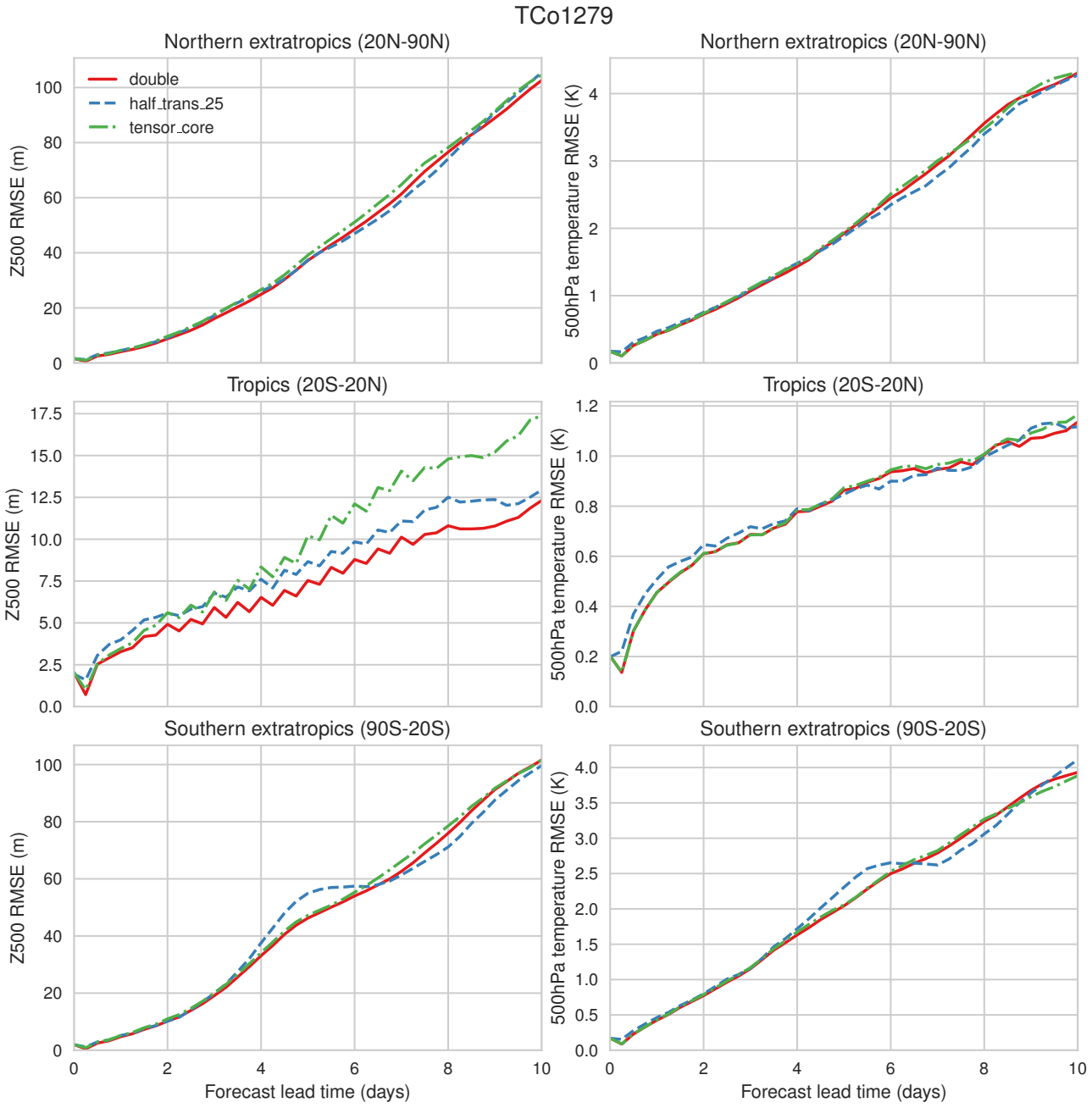

Figure 6: The root-mean-square error of the 500 hPa geopotential height field for various experimental setups verified with respect to analysis at TCo1279 resolution. Error metrics are averaged across multiple dates. Six dates were used for double and tensor_core whereas three dates were used for half_trans_25.

double and tensor_core setups. The tensor_core setup gives almost exactly the same forecast as the double setup. It is reassuring that there is minimal difference in the forecast skill of the double and tensor_core setups for specific extreme weather events as well as for large-scale spatially averaged verification metrics.

5 COMPUTATIONAL COST ESTIMATES

The computational cost savings of the half_trans and tensor_core setups with respect to the double-precision reference are difficult to estimate. The exact cost saving will depend on how the target architecture is configured, including the CPU-GPU connection and

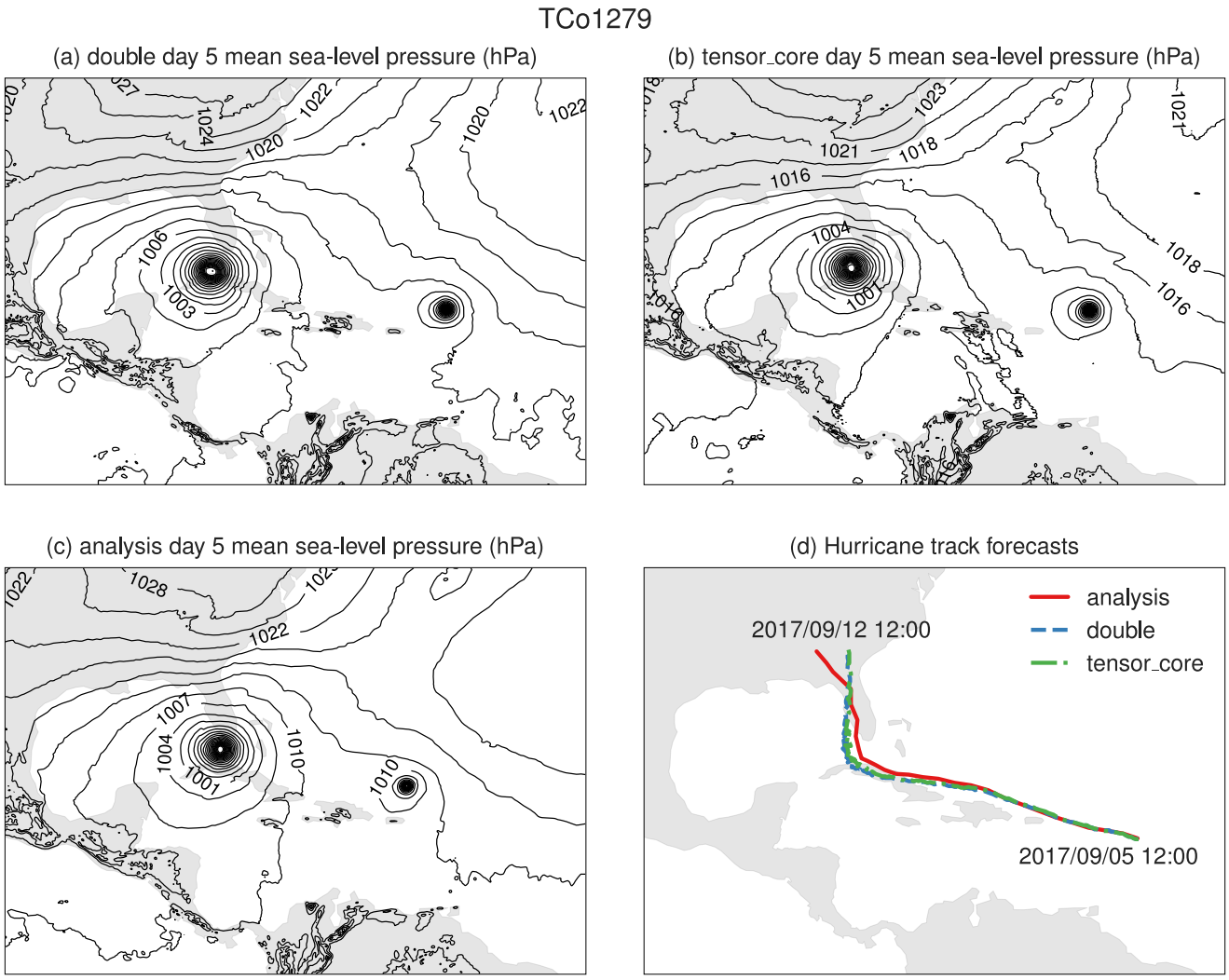

Figure 7: Forecasts of Hurricane Irma compared with analysis at TCo1279 resolution. Panes (a) and (b) show the mean sea-level pressure forecasts produced with the double and tensor_core setups, respectively, at 5 days lead time and pane (c) shows the corresponding analysis (actual conditions). Pane (d) compares the hurricane tracks up to day 7.

the node-node interconnect. However, we can provide a simple cost saving estimate from a computational complexity argument, which should apply to a single-node system. We present results for the symmetric part of the inverse Legendre transform, since the complexity is almost identical for the direct and antisymmetric parts.

We first consider the half_trans setup. By computing the total number of floating-point operations (FLOPs) performed at half- and double-precision, weighting the half-precision FLOPs appropriately, then dividing by the total number of FLOPs, we can obtain the cost of half_trans with respect to the fully double-precision setup. The number of FLOPs performed for a particular wavenumber m, assuming the standard "schoolbook" matrix-matrix multiplication algorithm, is given to a good approximation by

$$N_{\text{FLOPs}}(m) = N_{\text{fields}}(m) \times N_{\text{total wavenumbers}}(m) \times N_{\text{latitudes}}(m), \quad (7)$$

where $N_{\text{fields}}(m)$, $N_{\text{total wavenumbers}}(m)$ and $N_{\text{latitudes}}(m)$ are the number of fields, total wavenumbers and latitudes that the transform is evaluated over for this m, respectively. The number of fields is the same for all m and so can be factored out, apart from the factor of 2 that it introduces for all $m \neq 0$ wavenumbers (which have real *and* imaginary components, unlike the $m = 0$ wavenumber). The number of total wavenumbers decreases linearly as we increase m according to the triangular truncation paradigm. Finally, the number of latitudes also decreases as we increase m, as the number of longitudes decreases as we move from equator to pole for the cubic octahedral grid. The high m waves therefore cannot be represented at these high latitudes.

To compute the cost of the half_trans setup with respect to the double-precision reference, for a particular cutoff wavenumber c,

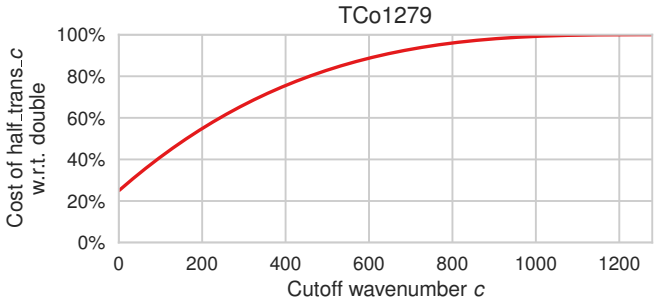

Figure 8: The cost of the half_trans setup with respect to the double-precision reference, as a function of the cutoff wavenumber c. Transforms use double-precision (half-precision) below (above) the cutoff wavenumber. Half-precision floating-point operations (FLOPs) are assumed to be four times cheaper than double-precision FLOPs.

we use the following formula:

$$C_{\text{half\_trans}}(c) = \frac{\sum_{m=0}^{c} N_{\text{FLOPs}}(m) + \sum_{m=c}^{N} S_{\text{half}} N_{\text{FLOPs}}(m)}{\sum_{m=0}^{N} N_{\text{FLOPs}}(m)}, \quad (8)$$

where S_{half} is the cost of a half-precision FLOP relative to a double-precision FLOP. We show this result as a function of the cutoff wavenumber c in Figure 8 for simulations with TCo1279 resolution. In this result, we assumed that a half-precision FLOP is four times cheaper than a double-precision FLOP, i.e. $S_{\text{half}} = 0.25$. If the cutoff wavenumber is N then all wavenumbers are computed at double-precision and there is no cost saving. If the cutoff wavenumber is 0 then all wavenumbers are computed at half-precision and the cost is 25% that of the double-precision reference. However, even when the cutoff wavenumber is 25 so that the first 25 wavenumbers are computed with double-precision (as in the previous experiments) the cost only increases to 29%.

A similar argument can be applied to the tensor_core setup, although in this case the cutoff wavenumber is just 1. Taking the quoted speed-up factor of 16 times from NVIDIA, the tensor_core setup should cost around 7% that of the double-precision setup. We stress again that this result will likely change for a multi-node system, and that there may be issues with load balancing if the compution of different zonal wavenumbers is spread across multiple nodes.

6 CONCLUSION

Here we have assessed the impact of reduced-precision floating-point arithmetic within an expensive portion of an atmospheric model by emulating hardware designed for deep-learning. Specifically, we considered the use of half-precision matrix-matrix multiplication and the NVIDIA Tensor Core for accelerating the Legendre transforms of the operational global weather forecasting model of ECMWF. Firstly, we have found that half-precision arithmetic can be used without causing model instabilities or crashes by using a simple rescaling technique, as long as the most sensitive 5% of the calculations are kept at double-precision. Secondly we have shown that the use of half-precision or the Tensor Core does degrade the forecast skill. However, the degradation is not significant when compared with other error sources that are present in real weather forecasts. We have shown that, in a probabilistic forecasting context with initial condition and model error, using half-precision or the Tensor Core gives a forecast competitive or equivalent to using double-precision. We have also demonstrated this result in a high-resolution, deterministic forecasting context. Finally, we have presented a simple cost estimate for the half-precision and Tensor Core Legendre transforms, though the exact numbers depend on the as-yet-unknown target architecture.

By 2025, ECMWF plan to increase the horizontal resolution of their ensemble forecasting system from TCo639 (roughly 18 km) to TCo1999 (roughly 5 km). To what extent can the positive results shown here be extended to higher resolutions? As horizontal resolution increases, the summing dimension of the matrix-matrix multiplication in both the direct and inverse Legendre transforms also increases. Therefore, the accumulated rounding error in this operation should grow also, more so when using half-precision than when using single-precision. However, it is extremely difficult to say *a priori* how this will impact a forecast product like the 10 day probabilistic forecast skill of 500 hPa geopotential height, given how errors feed back and interact in a nonlinear fashion within the model. Nevertheless, we are confident that a scale-selective approach, whereby larger scales (in our case represented by lower zonal wavenumbers) are kept at a higher precision than smaller scales, will enable the use of low-precision Legendre transforms at least up to TCo1999 and potentially beyond.

ACKNOWLEDGMENTS

The authors would like to thank Paul Dando at ECMWF for his help in running the IFS experiments and Vishal Mehta at NVIDIA for his help in developing the Tensor Core emulator.

Sam Hatfield is funded by the Natural Environment Research Council under grant number NE/L002612/1. Matthew Chantry was supported by a grant from the Office of Naval Research Global. Peter 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 programme under grant agreement number 675191. Tim Palmer received funding from the European Research Council (grant agreement number 741112) under the European Union's Horizon 2020 research and innovation programme.

REFERENCES

[1] Peter Bauer, Alan Thorpe, and Gilbert Brunet. 2015. The quiet revolution of numerical weather prediction. *Nature* 525, 7567 (2015), 47–55. https://doi.org/10.1038/nature14956

[2] Matthew Chantry, Tobias Thornes, Tim Palmer, and Peter Düben. 2019. Scale-Selective Precision for Weather and Climate Forecasting. *Monthly Weather Review* 147, 2 (2019), 645–655. https://doi.org/10.1175/MWR-D-18-0308.1

[3] Andrew Dawson and Peter D. Düben. 2017. Rpe v5: An emulator for reduced floating-point precision in large numerical simulations. *Geoscientific Model Development* 10, 6 (2017), 2221–2230. https://doi.org/10.5194/gmd-10-2221-2017

[4] Peter D. Düben, Hugh McNamara, and T. N. Palmer. 2014. The use of imprecise processing to improve accuracy in weather & climate prediction. *J. Comput. Phys.* 271 (2014), 2–18. https://doi.org/10.1016/j.jcp.2013.10.042

[5] Oliver Fuhrer, Tarun Chadha, Torsten Hoefler, Grzegorz Kwasniewski, Xavier Lapillonne, David Leutwyler, Daniel Lüthi, Carlos Osuna, Christoph Schär, Thomas C. Schulthess, and Hannes Vogt. 2018. Near-global climate simulation at 1km resolution: Establishing a performance baseline on 4888 GPUs

with COSMO 5.0. *Geoscientific Model Development* 11, 4 (2018), 1665–1681. https://doi.org/10.5194/gmd-11-1665-2018

[6] Sam Hatfield, Peter Düben, Matthew Chantry, Keiichi Kondo, Takemasa Miyoshi, and Tim Palmer. 2018. Choosing the optimal numerical precision for data assimilation in the presence of model error. *Journal of Advances in Modeling Earth Systems* (2018). https://doi.org/10.1029/2018MS001341

[7] Sam Hatfield, Aneesh Subramanian, Tim Palmer, and Peter Düben. 2018. Improving weather forecast skill through reduced precision data assimilation. *Monthly Weather Review* 146 (2018), 49–62. https://doi.org/10.1175/MWR-D-17-0132.1

[8] Nicholas J Higham. 2002. *Accuracy and Stability of Numerical Algorithms*. 1—-663 pages. https://doi.org/10.2307/2669725 arXiv:arXiv:1011.1669v3

[9] Bryan N. Lawrence, Michael Rezny, Reinhard Budich, Peter Bauer, Jörg Behrens, Mick Carter, Willem Deconinck, Rupert Ford, Christopher Maynard, Steven Mullerworth, Carlos Osuna, Andrew Porter, Kim Serradell, Sophie Valcke, Nils Wedi, and Simon Wilson. 2017. Crossing the Chasm: How to develop weather and climate models for next generation computers? *Geoscientific Model Development Discussions* September (2017), 1–36. https://doi.org/10.5194/gmd-2017-186

[10] Martin Leutbecher. 2018. Ensemble size: How suboptimal is less than infinity? *Quarterly Journal of the Royal Meteorological Society* June (2018), 1–22. https://doi.org/10.1002/qj.3387

[11] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. 2017. Mixed Precision Training. *CoRR* abs/1710.0 (2017). arXiv:1710.03740 http://arxiv.org/abs/1710.03740

[12] Andreas Müller, Willem Deconinck, Christian Kühnlein, Gianmarco Mengaldo, Michael Lange, Nils Wedi, Peter Bauer, Piotr K. Smolarkiewicz, Michail Diamantakis, Sarah-Jane Lock, Mats Hamrud, Sami Saarinen, George Mozdzynski, Daniel Thiemert, Michael Glinton, Pierre Bénard, Fabrice Voitus, Charles Colavolpe, Philippe Marguinaud, Yongjun Zheng, Joris Van Bever, Daan Degrauwe, Geert Smet, Piet Termonia, Kristian P. Nielsen, Bent H. Sass, Jacob W. Poulsen, Per Berg, Carlos Osuna, Oliver Fuhrer, Valentin Clement, Michael Baldauf, Mike Gillard, Joanna Szmelter, Enda O'Brien, Alastair McKinstry, Oisín Robinson, Parijat Shukla, Michael Lysaght, Michał Kulczewski, Milosz Ciznicki, Wojciech Piątek, Sebastian Ciesielski, Marek Błażewicz, Krzysztof Kurowski, Marcin Procyk, Pawel Spychala, Bartosz Bosak, Zbigniew Piotrowski, Andrzej Wyszogrodzki, Erwan Raffin, Cyril Mazauric, David Guibert, Louis Douriez, Xavier Vigouroux, Alan Gray, Peter Messmer, Alexander J. Macfaden, and Nick New. 2019. The ESCAPE project: Energy-efficient Scalable Algorithms for Weather Prediction at Exascale. *Geoscientific Model Development Discussions* January (2019), 1–50. https://doi.org/10.5194/gmd-2018-304

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

[14] T. N. Palmer, R. Buizza, F. Doblas-Reyes, T. Jung, Martin Leutbecher, G. Shutts, M. Steinheimer, and Antje Weisheimer. 2009. Stochastic Parametrization and Model Uncertainty. *ECMWF Tech. Memo. 598* (2009), 42. https://www2.physics.ox.ac.uk/sites/default/files/2011-08-15/techmemo598{_}stochphys{_}2009{_}pdf{_}50419.pdf

[15] A. J. Simmons, D. M. Burridge, M. Jarraud, C. Girard, and W. Wergen. 1989. The ECMWF medium-range prediction models development of the numerical formulations and the impact of increased resolution. *Meteorology and Atmospheric Physics* 40, 1-3 (1989), 28–60. https://doi.org/10.1007/BF01027467

[16] Tobias Thornes, Peter Düben, and Tim Palmer. 2017. On the use of scale-dependent precision in Earth System modelling. *Quarterly Journal of the Royal Meteorological Society* 143, 703 (2017), 897–908. https://doi.org/10.1002/qj.2974

[17] N. P. Wedi. 2014. Increasing horizontal resolution in numerical weather prediction and climate simulations: illusion or panacea? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 372, 2018 (2014), 20130289–20130289. https://doi.org/10.1098/rsta.2013.0289

[18] Nils P. Wedi, Mats Hamrud, and George Mozdzynski. 2013. A Fast Spherical Harmonics Transform for Global NWP and Climate Models. *Monthly Weather Review* 141, 10 (2013), 3450–3461. https://doi.org/10.1175/MWR-D-13-00016.1

[19] Hisashi Yashiro, Masaaki Terai, Ryuji Yoshida, Shin-ichi Iga, Kazuo Minami, and Hirofumi Tomita. 2016. Performance Analysis and Optimization of Nonhydrostatic ICosahedral Atmospheric Model (NICAM) on the K Computer and TSUBAME2.5. In *Proceedings of the Platform for Advanced Scientific Computing Conference*. ACM Press, New York, New York, USA, 1–8. https://doi.org/10.1145/2929908.2929911