PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A

rsta.royalsocietypublishing.org

Research

Cite this article: Wedi NP. 2014 Increasing horizontal resolution in numerical weather prediction and climate simulations: illusion or panacea? *Phil. Trans. R. Soc. A* **372**: 20130289. http://dx.doi.org/10.1098/rsta.2013.0289

One contribution of 14 to a Theme Issue 'Stochastic modelling and energy-efficient computing for weather and climate prediction'.

Subject Areas:
meteorology, applied mathematics, atmospheric science, computer modelling and simulation

Keywords:
spectral transform, numerical weather prediction, ensemble forecast, supercomputing

Author for correspondence:
Nils P. Wedi
e-mail: wedi@ecmwf.int

Increasing horizontal resolution in numerical weather prediction and climate simulations: illusion or panacea?

Nils P. Wedi

ECMWF, Shinfield Road, Reading RG2 9AX, UK

The steady path of doubling the global horizontal resolution approximately every 8 years in numerical weather prediction (NWP) at the European Centre for Medium Range Weather Forecasts may be substantially altered with emerging novel computing architectures. It coincides with the need to appropriately address and determine forecast uncertainty with increasing resolution, in particular, when convective-scale motions start to be resolved. Blunt increases in the model resolution will quickly become unaffordable and may not lead to improved NWP forecasts. Consequently, there is a need to accordingly adjust proven numerical techniques. An informed decision on the modelling strategy for harnessing exascale, massively parallel computing power thus also requires a deeper understanding of the sensitivity to uncertainty—for each part of the model—and ultimately a deeper understanding of multi-scale interactions in the atmosphere and their numerical realization in ultra-high-resolution NWP and climate simulations. This paper explores opportunities for substantial increases in the forecast efficiency by judicious adjustment of the formal accuracy or relative resolution in the spectral and physical space. One path is to reduce the formal accuracy by which the spectral transforms are computed. The other pathway explores the importance of the ratio used for the horizontal resolution in gridpoint space versus wavenumbers in spectral space. This is relevant for both high-resolution simulations as well as ensemble-based uncertainty estimation.

Royal Society Publishing

1. Introduction

Numerical weather prediction (NWP) requires an answer in real time with a window of approximately 1 h to run a medium-range global forecast that can be delivered in time to its customers. Increasingly, there is a need to address not only the primary question of the forecast outcome, but also to quantify the uncertainty with which the forecast is made. Today, ensembles of simulations with suitable perturbations are run to provide such an uncertainty estimate. Moreover, while computational efficiency remains one of the most pressing needs of NWP, there is an open question about how to make the most effective use of the affordable computer power that will be available over the next decades, while seeking the most accurate forecast possible. With increased computing capacity and corresponding advances in the numerical techniques applied (e.g. semi-implicit timestepping [1] and semi-Lagrangian advection [2]), there has been a steady increase in horizontal resolution, by approximately doubling the global horizontal resolution every 8 years at the European Centre for Medium Range Weather Forecasts (ECMWF). This rate reflects corresponding increases in computing power and provides the basis for an increase in the time range for which successful forecasts can be made by 0.5–1 day per decade. However, this rate is too slow for the growing demand to run subkilometre scale, convection-resolving global weather and climate simulations within the next decade.

At present, the factors driving continued horizontal resolution increases are (i) at current resolutions, important processes determining the vertical redistribution of energy in the atmosphere are not resolved, (ii) more accurate resolved representations of the forcing, i.e. topography, vegetation, land-use fields and ocean currents, have a decisive impact on the atmospheric dynamics, (iii) so far horizontal resolution increases have improved the skill of NWP and climate predictions, and (iv) larger problems scale better on massively parallel platforms. In the future, however, model development will be constrained by other drivers: these are, from a technical point of view, the energy efficiency and the (hardware-related) reliability of massively parallel computations, and from a scientific point of view, the reliability of forecasts together with a quantitative assessment of the uncertainty.

Uncertainty increases with increased degrees of freedom. Global NWP has reached the threshold of permitting and resolving convection explicitly, where the convective-scale uncertainty and thus errors in the forecast may grow upscale more quickly, and accounting for various sources of uncertainty in simulations that explicitly simulate moist convection is an active area of research (see [3] and references therein). It is also unclear how this will alter predictability at the larger scales. Therefore, there is a need to appropriately address forecast uncertainty with increasing resolution and how to best represent this uncertainty in the model. There are different sources of uncertainty to be considered, uncertainty around the best estimate of the initial state, uncertainty arising from imperfect model assumptions, and uncertainty arising from the choices made for particular numerical algorithms associated with numerical truncation errors as well as arithmetical round-off errors. An informed decision on the modelling strategy for harnessing exascale, massively parallel computing power thus also requires a deeper understanding of the model's sensitivity to uncertainty (and possibly to the lack of hardware reliability)—for each part of the model—and ultimately a deeper understanding of multi-scale interactions in the atmosphere and their numerical realization in ultra-high-resolution NWP and climate simulations.

In the Integrated Forecasting System (IFS) at ECMWF, horizontal resolution is expressed by the cut-off spectral truncation number N of the spherical harmonics series expansion of the prognostic variables. The spectral transform method has been successfully applied at ECMWF for approximately 30 years, with the first spectral model introduced into operations in April 1983. Spectral transforms on the sphere involve discrete spherical harmonics transformations between physical (gridpoint) space and spectral (spherical harmonics) space. The spectral transform method was introduced to NWP following the work of Eliasen *et al*. [4] and Orszag [5], who pioneered the efficiency obtained by partitioning the computations. One part of the computations is performed in physical space, where products of terms, the semi-Lagrangian or Eulerian

advection, and the physical parametrizations are computed. The other part is solved in spectral space, where the Helmholtz equation arising from the semi-implicit timestepping scheme can be solved easily and horizontal gradients on the (reduced) Gaussian grid are computed accurately, particularly the Laplacian operator that is so fundamental to the propagation of atmospheric waves. The success of the spectral transform method in NWP in comparison with alternative methods has been overwhelming, with many operational forecast centres having made the spectral transform their method of choice, as comprehensively reviewed in [6].

A spherical harmonics transform is a Fourier transformation in longitude and a Legendre transformation in latitude, thus keeping a latitude–longitude structure in gridpoint space. The Fourier transform part of a spherical harmonics transform is computed numerically very efficiently by using the fast Fourier transform [7,8] that reduces the computational complexity to $\propto \mathcal{O}(N^2 \log N)$. However, the conventional Legendre transform has a computational complexity $\propto \mathcal{O}(N^3)$, and with increasing horizontal resolution the Legendre transform will eventually become the most expensive part of the computations in terms of the number of floating point operations and subsequently the elapsed (wall-clock) time required. Owing to the relative cost increase of the Legendre transforms compared with the gridpoint computations, very high-resolution spectral models were believed to eventually become prohibitively expensive. However, the recent implementation of a fast Legendre transform (FLT) [9] mitigates the concern about the disproportionally growing computational cost. FLTs are a paradigm algorithm for trading formal accuracy for computational efficiency, following the seminal work of Tygert [10,11], the algorithm for the rapid evaluation of special functions described in O'Neil *et al*. [12] and the efficient interpolative decomposition matrix compression technique described in Cheng *et al*. [13].

One issue that arises with the new emerging computing architectures is the mandatory restriction of data movement across physically distant processors owing to its high cost in terms of energy and wall-clock time. In this paper, the IFS at ECMWF is used to illustrate some of the issues facing operational NWP while contributing to a better understanding of future horizontal resolution increases, and where formal accuracy may be traded for enhanced numerical efficiency. For example, in IFS, the only mechanism to compute horizontal derivatives is via the global spectral transforms. Hence, if sufficiently accurate local derivative computations can be constructed outside spectral space, the remaining use of the spectral representation would be the solution of the semi-implicit linear problem (the nonlinear residual is computed explicitly in gridpoint space) including the trivial representation of the horizontal Laplacian of a variable, $\nabla^2\psi \equiv -(n(n+1)/a^2)\psi$, where ψ is a spectral coefficient, a is the Earth's radius and n is the total wavenumber. The choice of the cut-off truncation wavenumber N is dictated by accurately ('alias-free') representing linear or quadratic terms. Historically, this led to choosing the number of longitudes along a given Gaussian latitude greater or equal to $2N+1$, and the number of latitudes $\geq(2N+1)/2$ for the so-called *linear grid* [14], whereas $\geq 3N+1$ longitudes with $\geq(3N+1)/2$ latitudes for the so-called *quadratic grid* [4,15,16]. In addition, a reduced grid is introduced [17], where the number of longitudes is reduced towards the poles, keeping the relative distances between points approximately constant, i.e. quasi-uniform. The reduction of the gridpoints with increasing geographical latitude θ away from the equator follows the rule with $3N_{\mathrm{r}}+1$ gridpoints for the reduced number of waves N_{r} (see [18] and also §3 for more details). Notably, in IFS, the number of Fourier modes on each latitude towards the poles is further optimized, reducing proportional to $1/(2+\cos^2(\theta))$, leading effectively to a linear grid for high latitudes with any truncation choice. Thus, another important role of the spectral transform method emerges, the process of 'filtering', and specifically on the sphere, where the transformation to spectral space provides for an 'ideal' polar filter.

Given the relatively larger cost increase of the spectral computations with increasing resolution, and the Legendre transforms in particular, there may be a case for solving the linear global semi-implicit problem more cheaply by using fewer wavenumbers than dictated by the linear grid choice. This hypothesis is explored in this paper by comparing the numerical efficiency and the large-scale hemispheric meteorological forecast accuracy—the ultimate measure of

success in NWP—for the linear, the quadratic and a newly formulated cubic grid representation (as defined below) at very high resolution. Notably, the advection, right-hand-side forcing terms, the subgrid-scale effects and also the perturbations used in ensemble forecasts are computed in gridpoint space that would continue to increase in resolution, whereas the spectral truncation remains constant or is increased more slowly. It is of particular interest how these choices influence the effective resolution of the model.

The article is organized as follows. Section 2 describes experiments with FLTs using various degrees of approximation in the FLT algorithm, while measuring the acceleration of the computations and comparing the meteorological results. These results extend the analysis provided in Wedi *et al*. [9]. Section 3 explores new options for the duality of gridpoint and spectral space computations and their respective influence on efficacy. Finally, §4 draws some conclusions.

2. Experimenting with the fast spherical harmonics transform

Because the fastest numerical methods used in geophysical fluid dynamics scale linearly with the number of gridpoints (i.e. proportional to N^2), the cost of the Legendre transforms would not be competitive at problem sizes with $N = O(1000)$, and very high-resolution spectral models may become prohibitively expensive. However, up to a resolution of approximately $N = 2047$, the very high rate of floating point operations per second (flops) achieved in matrix–matrix multiplications used within the spectral computations masks the $\propto N^3$ cost of this part of the IFS model. For higher resolutions, the implementation of the FLTs into the spectral transform model IFS has been shown to mitigate the increased computational cost by scaling according to $\mathcal{O}(N^2 \log^3 N)$ [9].

FLTs represent a paradigm algorithm for trading formal accuracy and computational efficiency. The essence of the so-called butterfly algorithm [11,12] is that matrices arising in part of the summations may be compressed using an interpolative decomposition [13], such that

$$|\mathcal{S}_{r\times s} - C_{r\times k} A_{k\times s}| \leq \epsilon, \tag{2.1}$$

where matrix $C_{r\times k}$ constitutes a subset of the columns of matrix $\mathcal{S}_{r\times s}$ and where matrix $A_{k\times s}$ contains a $k \times k$ identity matrix with k being called the ϵ-rank of submatrix $\mathcal{S}_{r\times s}$ [13,19]. The parameter ϵ defines the accuracy required in the compression part of the algorithm. Wedi *et al*. [9] find that with $\epsilon = 10^{-7}$ equivalent meteorological accuracy as measured in terms of hemispheric root-mean square error (RMS) and anomaly correlation of 500 hPa geopotential height and other parameters (not shown)—typically used to verify technical model changes—is obtained. Further to the analysis presented in that paper, here we analyse more closely the effect of compression accuracy and the impact on the efficiency of the computations. Table 1 summarizes these results. Flop refers to the counted number of floating point operations used in a 48 h forecast for inverse and direct transforms, respectively. All results shown in table 1 with the specified choices for ϵ have an equivalent meteorological performance with the same RMS as defined above when averaged over seven independent selected dates. A stronger compression does reduce the computational cost further, but this does impair ultimately the meteorological forecast. We find that while the number of floating point operations continues to reduce with successively lower thresholds of ϵ, the dominant saving is already achieved with $\epsilon = 10^{-10}$. Notably, we do not find any degradation in the meteorological result nor in global kinetic energy spectra (not shown) with $\epsilon = 10^{-4}$ for the inverse transform only. Incidentally, the inverse transform is the most costly part, because, in this step, the derivatives also are computed. By contrast, the direct transforms are sensitive to the choice $\epsilon < 10^{-7}$. In [9], the model was shown to be numerically unstable after 5 days of simulation with $\epsilon = 10^{-2}$, showing a build-up of energy at the tail of the kinetic energy spectrum. This instability can be eliminated if the $\epsilon = 10^{-2}$ threshold is only used for the inverse transform.

3. Gridpoint versus spectral computations

For ensemble-based uncertainty estimation in operational forecast centres, the execution speed at which the ensemble members can be computed is critical. Often the individual forecast quality

Table 1. Results for forecasts using FLTs with different compression ϵ (split into *inverse* and *direct* transform; see text for details).

truncation N	FLT	ϵ_{inv}	ϵ_{dir}	flop_{inv} ($\times 10^7$)	flop_{dir} ($\times 10^7$)
1279	no	—	—	46.4	33.7
1279	yes	10^{-10}	10^{-10}	36.6	26.3
1279	yes	10^{-4}	10^{-7}	34.2	24.6
2047	no	—	—	249.6	181.5
2047	yes	10^{-7}	10^{-7}	153.3	110.5
2047	yes	10^{-4}	10^{-7}	147.1	110.5

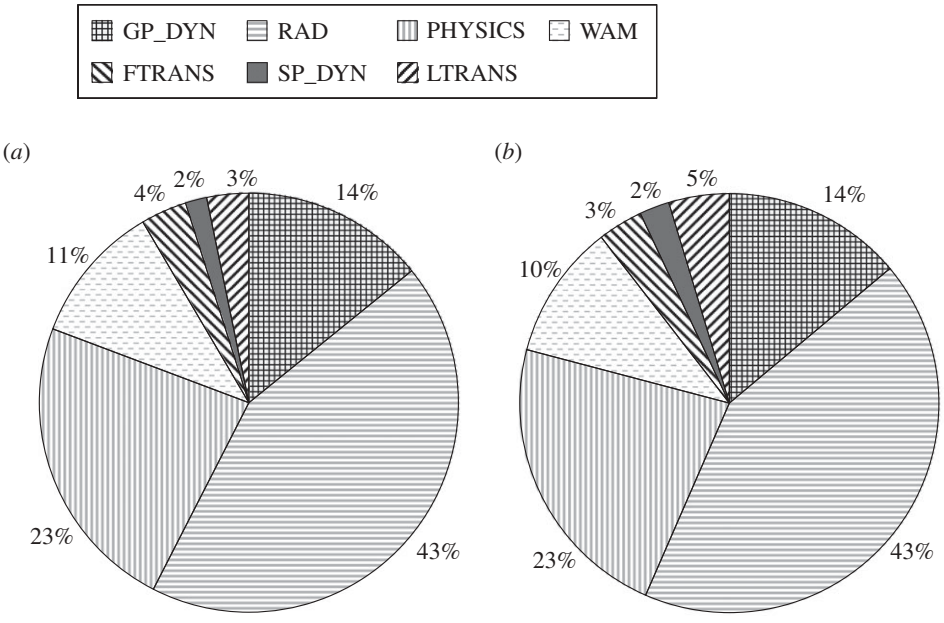

Figure 1. Cost distribution of a 10 day forecast at T_q1364 resolution (*a*) and the cost distribution at T_l2047 (*b*). Both forecasts use the same number of gridpoints. The computations associated with spectral space ($\approx$10% of the total), including the transpositions from gridpoint to spectral to gridpoint, are *FTRANS* (Fourier transforms), *LTRANS* (Legendre transforms) and *SP_DYN* (semi-implicit spectral computations). *GP_DYN* represents the semi-Lagrangian gridpoint computations (14%), *RAD* are the radiation gridpoint computations (43%), *PHYSICS* represents the other physical parametrization calculations (23%), and *WAM* is the cost of the ocean surface wave model. Although not visible in the percentages here, the cost of the spectral computations is reduced by approximately 30% for the T_q1364 quadratic grid.

is degraded as a result, and the ensemble spread is larger. This effect compensates sometimes for a tendency to have too little spread, or in other words, an over-confident forecast. On the other hand, improving the quality of the forecast model for each individual ensemble member leads to a sharper probability distribution and higher confidence in the forecast, accompanied by a smaller spread. In order to improve the efficiency, while maintaining a high level of accuracy, it is important to know where the computational effort is spent and where it should be spent. Figure 1 illustrates the computational cost distribution of two simulations with the equivalent number of gridpoints. Figure 1 refers to simulations with $N = 1364$ (figure 1*a*) and $N = 2047$ (figure 1*b*) spectral wavenumbers. The latter simulation is more expensive in the spectral part of the computations by approximately 30%. Notably, about 60% of the cost is spent in gridpoint space calculations with the largest part in the physics and radiation calculations. The simulations

were coupled to a 0.1 degree wave model which is approximately doubling the resolution (and the relative computational effort) compared with the operational configuration. Moreover, the cost distribution is representative of a timestep when the radiation calculations are called. Typically, the cost of the radiation is reduced by reducing the frequency (in time) of these calculations and by reducing the grid on which these calculations are performed (a coarser grid corresponding to T_l799 has been used in all simulations in this paper). However, both choices impact negatively the meteorological performance, especially near coastlines with sharp gradients in radiative properties and where the differences between different resolution grids are very apparent. The T_q1364 refers to a quadratic grid simulation, where aliasing in the quadratic terms is avoided by ensuring that the number of gridpoints used along equatorial Gaussian latitudes is not less than $3N+1$. If aliasing in the terms involving triple products is to be avoided, the number of gridpoints used should not be less than $4N+1$, leading to a cubic grid. The T_l2047 refers to a linear grid simulation where the number of gridpoints used along equatorial Gaussian latitudes is $2N+1$, thus admitting aliasing in quadratic and higher-order terms. For practical reasons, the number of points along the Gaussian latitudes in the reduced grid (approx. outside of ± 30 degrees latitude) are always selected according to the $3N_r+1$ rule [18]. Consequently, east–west aliasing for cubic terms in the cubic grid remains, and for the linear grid, east–west aliasing of quadratic terms is removed in this way.

The linear grid has been used at ECMWF for decades [15] as the remaining aliasing could be controlled by other means, such as horizontal diffusion or special de-aliasing filters. In return, the higher spectral resolution offered substantial advantages. Most importantly, at moderate horizontal resolutions, the orographic forcing, the representation of the meridional derivatives, and the Laplacian operator could be enhanced in this way for little extra cost. However, at ultra-high resolution, we find that the situation is reversed. The relative cost increase of the Legendre computations combined with more costly de-aliasing procedures (see the difference in flop in table 1 between T_l2047 and T_l1279) suggests that the quadratic or cubic grid may be more efficient at higher resolutions. For example, at climate resolutions of T_l511 (or an equivalent horizontal grid spacing of $\Delta \approx 39$ km), the difference to the quadratic grid is 170 waves, which appears meteorologically significant—covering synoptic- and mesoscale—yet it is computationally insignificant. However, at the next planned operational resolution upgrade to T_l2047 ($\Delta \approx 10$ km), the difference is 683 waves, and at the next resolution doubling T_l3999 ($\Delta \approx 5$ km) the difference is 1333 waves. In both cases, the range of the last third of the spectrum covers smaller and smaller—arguably less predictable—scales, and is subject to aliasing and thus special filtering in the case of the linear grid. This raises the question if the additional expense for higher spectral resolution is significant or even appropriate at these very high resolutions. Recent evaluations of the effective resolution of NWP simulations suggest a range of 6–8Δ [20,21], which is above the filter range of the quadratic grid (2–3Δ) [16] and the cubic grid (3–4Δ).

On the other hand, we know from experimentation (not shown) that the representation of orography is important and that the additional wavenumbers improve both the forecast and the assimilation [15]. So far, the orography has been derived from a latitude–longitude representation at 1 km resolution by first applying a band pass filter with a physical filter width equivalent to the average distance of gridpoints in the target Gaussian grid (i.e. 16 km for the operational T_l1279 resolution). The resulting field (still on the original grid) is interpolated to the Gaussian grid and a direct spectral transform is applied. In spectral space, the field is filtered with a high-order spectral Butterworth-type filter to avoid aliasing at the smallest scales. Here, a new approach is adopted. In order to benefit from orography information at relatively higher resolution (despite a nominally lower truncation wavenumber for the quadratic grid), the underlying orography is derived from the $N=7999$ spectral representation of orography by spectral interpolation to $N=1364$ and subsequent inverse transform to the corresponding Gaussian grid. Figure 2 illustrates how the new approach retains more variance in the orography field compared with the orographies derived previously, with the dotted line (T_q1364 from T_l7999) above both the standard T_q1364 and the linear grid T_l2047 in the wavenumber range 700–1200. All simulations with the quadratic and

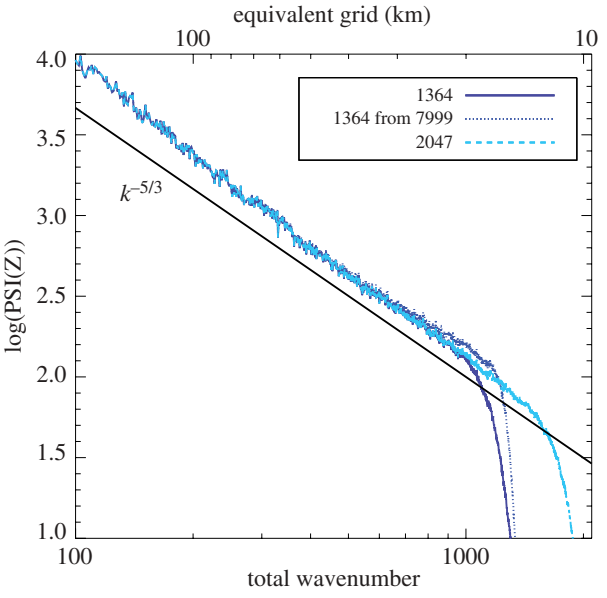

Figure 2. Comparison of global orographic variance (power density spectrum in m^2) at different horizontal resolutions. The orographies using the standard procedure described in the text at T_q1364 (solid) and T_l2047 (dashed) resolutions are compared with the T_q1364 (dotted) orography derived by spectral interpolation from the T_l7999 orography. (Online version in colour.)

cubic grid in the following are done with the newly derived orographic field. One might think that the same new approach could be used for the linear grid simulations such as T_l2047, but this is not the case. The higher variance conflicts with the de-aliasing filter used for the linear grid and it is much more difficult (if possible at all) to control the aliasing in this case. As a result, this leads to unrealistic spectral blocking at the end of the spectrum and worse forecast skill (not shown). The effect of the orographic forcing can also be seen in the global horizontal kinetic energy spectra derived from the vorticity and divergence at different vertical levels after 5 days of simulation. Figure 3 compares the spectra of the T_c1023 and the T_l2047 simulations with a higher resolution reference simulation at T_l3999. Figure 3*a* is near the surface and figure 3*b* at a model level at approximately 500 hPa. The low wavenumber range is identical and only the high wavenumber part is shown in figure 3. The $k^{-5/3}$ spectrum is clearly visible near the surface but less pronounced in the mid-troposphere. In both cases, the cubic grid T_c1023 evinces a higher effective resolution in the 300–900 wavenumber range, assuming that the T_l3999 represents the best estimate of the expected 'truth' and based on the experience that the high wavenumber part of the horizontal kinetic energy spectrum asymptotes towards higher amplitude with increasing horizontal resolution (see also fig. 5 in [9]). The increase in amplitude in the 300–900 wavenumber range despite coarser spectral truncation indicates the importance of the increase in resolution in gridpoint space. Comparing the T_q1364 and the T_l2047 with both using the same number of gridpoints (not shown), both appear very similar up to the point where de-aliasing filters and/or horizontal diffusion act (the last third of the spectrum in the linear grid case). This is in particular the case with the new way of deriving the orography. However, comparing upper tropospheric spectra of the T_c1023 and T_l2047 simulations (not shown), the slopes extend in the linear grid case beyond the cubic grid truncation along the same slope, suggesting more effective resolution for the higher (linear grid) wavenumber simulation.

The spectra would suggest that increasing the number of gridpoints is beneficial even if gridpoint-vertical-column calculated physical parametrizations are used to represent subgrid-scale diabatic forcings. On the contrary, not much may be gained for the larger-scale upper tropospheric motions by the use of the linear grid compared with the quadratic grid. But when comparing with the T_l3999, energetically significant differences can be seen at wavenumbers

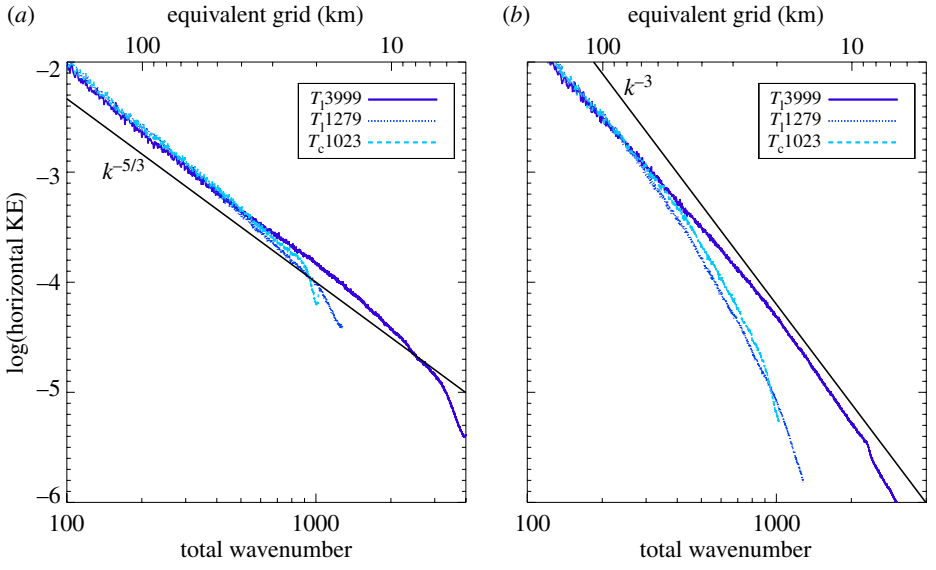

Figure 3. Horizontal kinetic energy (KE) spectra plots. Comparison of global spectra after 5 days of simulation for the resolutions T_c1023 and T_l1279 at the lowest model level ≈10 m height (*a*) and at a mid-tropospheric model level ≈ 500 hPa (*b*). The cubic grid T_c1023 evinces a higher effective resolution despite a lower spectral truncation, indicating the importance of the increase in resolution in gridpoint space and the revised derivation of the underlying orography. For reference, the spectra of a T_l3999 simulation after 5 days (albeit for a different date) are also shown. (Online version in colour.)

much less than the nominal resolution of the other simulations, implying a general need for continued resolution increases in the future. The picture is incomplete since figure 3 only shows the horizontal kinetic energy, and the differences could suggest a substantial repartition of vertical and horizontal energy that depends on the model resolution and how the physical parametrizations are able to emulate the subgrid scale effects driving this repartition and any potential upscale effects.

In the following, we investigate the sensitivity to increasing the horizontal resolution in the gridpoint space part of the computations only. Technically, in the experiments, the number of gridpoints are actually kept constant, while using successively lower spectral truncations. All simulations start with the same T_l1279 initial conditions and are conducted with 137 vertical levels. The different simulations are summarized in table 2 comparing the cost and effect. The number of gridpoints in the table refer to a single model level. All simulations ran on the IBM Power 7 supercomputer using 256 MPI tasks with each using 16 OMP threads, equalling 4096 compute tasks in each simulation. As illustrated in table 2, the robustness of the quadratic grid may be stretched further, by increasing the timestep of the simulation compared with the linear grid one. The size of the timestep plays a crucial role in the success of NWP and climate, where limits due to fast but energetically insignificant waves have been removed by semi-implicit timestepping, and where the advective Courant–Friedrichs–Lewy limit has been substantially enhanced by the semi-Lagrangian advection algorithm. We postulate that 'large timestep' solutions continue to be an important aspect of efficient NWP and climate integrations as long as the physically relevant and resolved time scales remain larger. In the following comparison, the T_l2047 simulations use a (experimentally determined) maximum permissible timestep $\Delta t = 450$ s, and two T_q1364 simulation series are done with timestep $\Delta t = 450$ s and $\Delta t = 600$ s, respectively. Equivalent control simulations have been done with the currently operational resolution T_l1279 using $\Delta t = 600$ s. The combination of the reduction in the truncation and the increase in the timestep size leads to a significant overall speedup in terms of achieved forecast days per day of 44%. Notably, the ratio of the simulated forecast days per real day (FCday) is reduced by slightly less than half when increasing the number of gridpoints by a

Table 2. Comparison of linear T_l1023, T_l1279, T_l2047, quadratic T_q1364 and cubic T_c1023 grid simulations on 4096 tasks on the IBM Power 7 (the lowest resolution grid simulation T_l1023 used 2048 tasks). Mean values for the anomaly correlation (acc Z500) of the 500 hPa geopotential surface and root-mean square error (RMS T850) of the 850 hPa temperature surface are representative for the Northern Hemisphere after 8 days. All data of these simulations are stored in the ECMWF Mars data archive.

truncation N	FC per day	Δt (s)	no. gridpoints	acc Z500 (%)	RMS T850 (K)
T_l1023	225	600	1 373 624	65.36	3.72
T_l1279	359	600	2 140 702	64.76	3.70
T_l2047	117	450	5 447 118	64.29	3.70
T_q1364	131	450	5 447 118	65.48	3.63
T_q1364	169	600	5 447 118	65.48	3.68
T_c1023	173	600	5 447 118	65.95	3.66
T_c1023	196	720	5 447 118	64.80	3.70

factor ≈2.5, while changing the spectral resolution from T_l1279 to T_q1364, indicating a near-linear scaling with the number of gridpoints and with the number of timesteps. In addition, two sets of simulations have been done with the newly generated T_c1023 cubic grid, using $\Delta t = 600$ s and $\Delta t = 720$ s, respectively. These results are also summarized in table 2 indicating an equivalent or better hemispheric meteorological performance and a further speedup in FCday. These may also be compared with the results of a T_l1023 linear grid, where the spectral truncation is kept constant but the number of gridpoints is approximately one-quarter of the cubic grid.

The meteorological accuracy is measured in terms of northern hemispheric RMS (figure 4*a–c*), and northern hemisphere anomaly correlation (acc) of the 500 hPa geopotential height surface (figure 4*d–f*) and other parameters (see also table 2). The data of all simulations have been truncated to the same $N = 120$ prior to calculating the scores. The results show two aspects. First, with the same high-resolution physical grid ($\Delta \approx 10$ km), we find neutral or improved results for the simulations with relatively coarser spectral truncation (quadratic or cubic) compared with the linear grid. Second, with a higher resolution physical grid, ($\Delta \approx 10$ km) compared with ($\Delta \approx 16$ km), and similar spectral truncation, T_c1023 and T_q1364 compared with T_l1279, we also find improved RMS and acc scores. The improvement in scores is further confirmed by comparing the T_c1023 with the T_l1023 simulations (see also table 2) with physical grids ($\Delta \approx 10$ km) and ($\Delta \approx 20$ km), respectively. Over the series of 27 selected forecast dates some values are reaching significance at the 95% level (bars above the zero line), despite the relatively short series. As can be seen from table 2, the skill is still high at day 8 with an average of approximately 65% anomaly correlation for the 500 hPa geopotential surface. Moreover, we find that near surface parameters such as the forecast significant wave height improve in the medium-range forecast with the quadratic and cubic grid simulations (not shown), moving closer to the verifying analysis. The series is perhaps too limited to conclude generally on the relative performance of either grid/resolution configuration, especially because all forecasts start from the same resolution initial conditions and do not include corresponding changes in the assimilation system. Nevertheless, for the purpose of this paper, it clearly underlines the message that trading some formal resolution and numerical accuracy at the smallest scales for computational and ultimately energy efficiency in ensemble-based uncertainty estimation is possible in various ways. Based on these results, increasing the resolution more rapidly in gridpoint space than in the corresponding spectral space emerges as a compelling concept for increasingly higher resolutions, in contrast to the practice in the past years. In gridpoint space, the forcings, the subgrid-scale effects and also the perturbations used in ensemble forecasts are computed, leading to relatively more spatial variability in these fields commensurate with the observations. Interestingly, the idea could also be used to initialize a cubic grid ensemble forecast with a high-resolution surface analysis without the need to interpolate the surface initial conditions, and thus eliminate the associated degradation due to interpolation.

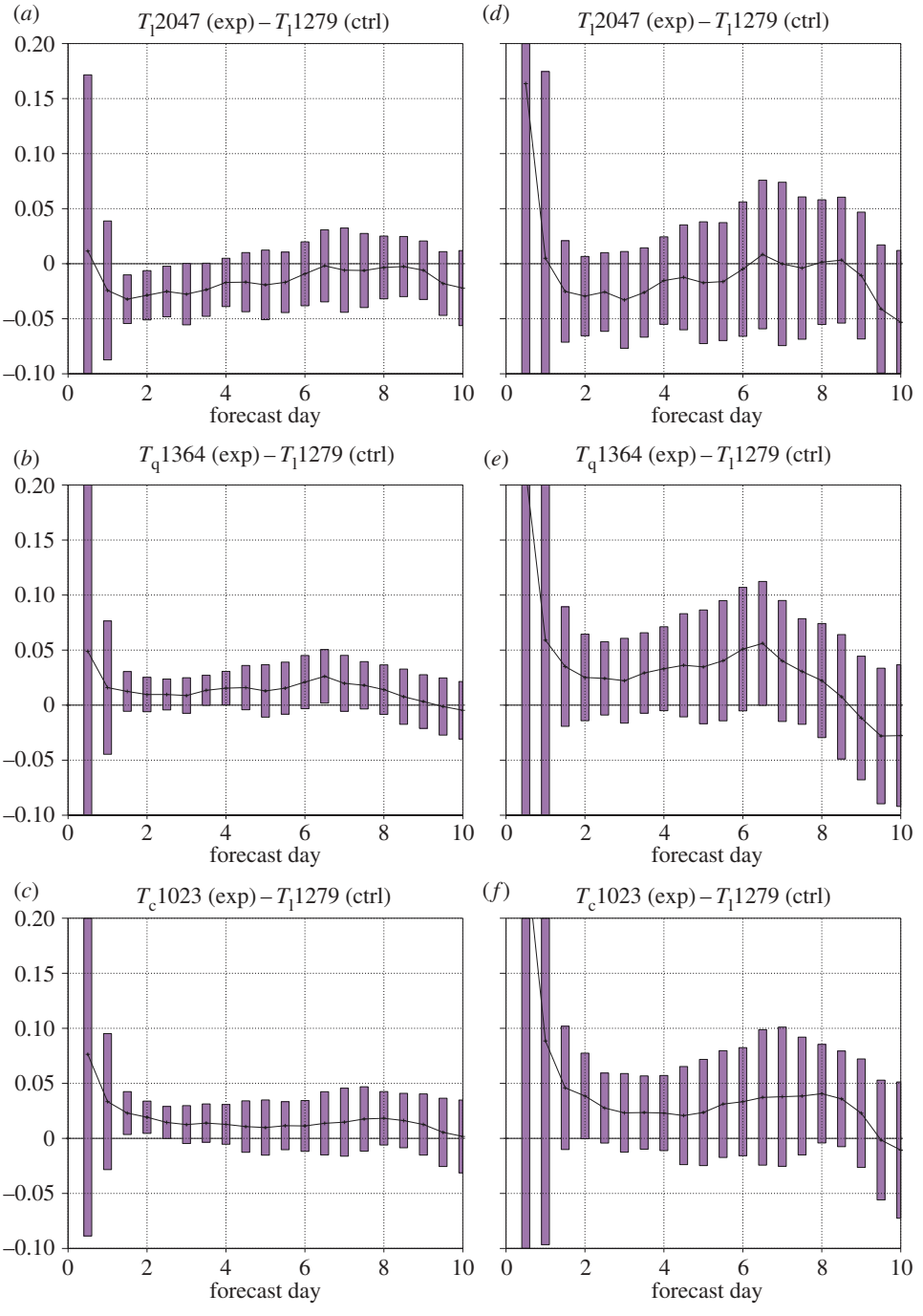

Figure 4. 500 hPa geopotential height root-mean square error (*a–c*) and anomaly correlation (*d–f*) in the Northern Hemisphere comparing the experiment (exp) T_l2047, T_q1364, and T_c1023 simulations against the control (ctrl) linear grid configuration T_l1279 for a series of 27 forecasts with different initial dates (every 15 days) for the period 20 August 2012–1 September 2013. The line graph shows the mean difference of experiment minus control normalized by the mean of both. The sign of the differences is chosen such that positive values always indicate that the experiment is better than the control. Confidence intervals (bars) are computed according to the *t*-test with a specified 95% confidence interval. (Online version in colour.)

4. Conclusion

ECMWF plans to implement a global horizontal resolution of approximately 10 km by 2015 for its assimilation and high-resolution forecasts, and approximately 20 km for the ensemble forecasts. The benefit of using the FLTs with enhanced compression in this resolution range is found to be limited. Moreover, the results show that the primary effect of reducing the computational effort is already achieved with a tiny non-zero ϵ. With further compression, the inverse transforms have been found to be less sensitive to a lower ϵ than the direct transforms. The efficiency gain in the floating point operations required by using the FLTs is substantial, but up to T_l2047 the gain in terms of wall-clock time is relatively small. The scales resolved in the simulations presented in this paper are still hydrostatic and questions remain about the importance of resolving the repartition of energy due to convective motions rather than parametrizing their effect in medium-range forecasts. It has been found that the efficiency and accuracy of the hydrostatic, semi-Lagrangian, semi-implicit solution procedure using the spectral transform method may be enhanced substantially by moving to the quadratic or cubic grid equivalents at these and higher resolutions. Notably, a substantial increase in efficiency of the cubic grid compared with the linear grid can be expected in the future from the corresponding reduction in the cost of transpositions and their associated parallel communications at larger processor counts. Especially, when combined with local calculations (e.g. of derivatives) in gridpoint space, the spectral semi-implicit solution procedure may be viewed as a small-scale filter, and an even further reduction of the wavenumbers involved (e.g. using a cubic or quartic grid) may be permissible as long as the relevant (large-scale) wave motions are sufficiently accurately captured. The T_c1023 simulations presented here are a first step in this direction with a smallest half-wavelength of 20 km in spectral space and a 10 km spacing in gridpoint space. For the high resolutions presented in this paper, with either keeping the grid constant or keeping the spectral truncation constant, the combination of a relatively coarser spectral truncation and a finer physical grid is meritorious. It is speculated that the mathematically correct filtering of aliased noise, the higher resolution in the computation of all nonlinear right-hand-side forcings and the relatively smaller physical distance of the interpolation stencil, associated with less damping in the semi-Lagrangian interpolations, contribute to this positive result. In addition, the parametrized physical forcings and surface interactions are calculated at relatively higher resolution. Notably, all moist quantities remain in higher resolution gridpoint space throughout the simulation. While Lander *et al.* [22] argued for a coarser 'physics' grid on the basis that parametrizations may be forced wrongly by poorly resolved flow features, their basic idea is not necessarily contrary to the results presented here. Horizontal divergence and thus the resolved vertical velocity ω are notably filtered to the coarser spectral truncation which in return provide feedback to the physical parametrizations, e.g. convection.

In conclusion, horizontal resolution increases in NWP and climate prediction are likely to continue to provide improvements in forecast quality and offer new opportunities for uncertainty estimation. However, blunt increases are not a *panacea* without adjusting the numerical techniques applied and are likely to be unaffordable or, worse, they may not lead to the desired improvements. To the contrary, the simulations may become *illusory* in that they provide solutions that appear more realistic, but potentially with impaired predictability. Thus, merely facilitating scalability through code adaptation is unlikely to be sufficient for successful future global NWP and climate predictions. Overall, the results presented here pose interesting new questions about the nature of linear and nonlinear multi-scale interactions in the atmosphere and how they are best represented and solved for in global simulations of weather and climate.

Acknowledgements. I thank Christian Kühnlein, George Mozdzynski and Peter Janssen for discussions and their helpful comments on an earlier version of the manuscript. Moreover, I acknowledge fruitful discussions with Piotr Smolarkiewicz, Martin Leutbecher and Peter Düben. Finally, I also thank the two anonymous reviewers for their comments and suggestions that improved the presentation.

References

1. Robert A, Henderson J, Turnbull C. 1972 An implicit time integration scheme for baroclinic models of the atmosphere. *Mon. Weather Rev.* **100**, 329–335. (doi:10.1175/1520-0493(1972)100<0329:AITISF>2.3.CO;2)
2. Ritchie H. 1988 Application of the semi-Lagrangian method to a spectral model of the shallow water equations. *Mon. Weather Rev.* **116**, 1587–1598. (doi:10.1175/1520-0493(1988)116<1587:AOTSLM>2.0.CO;2)
3. Kühnlein C, Keil C, Craig GC, Gebhardt C. In press. The impact of downscaled initial condition perturbations on convective-scale ensemble forecasts of precipitation. *Q. J. R. Meteorol. Soc.* (doi:10.1002/qj.2238)
4. Eliasen E, Machenhauer B, Rasmussen E. 1970 On a numerical method for integration of the hydrodynamical equations with a spectral representation of the horizontal fields. Report 2, Institut for Teoretisk Meteorologi, University of Copenhagen, Denmark.
5. Orszag SA. 1970 Transform method for calculation of vector coupled sums: application to the spectral form of the vorticity equation. *J. Atmos. Sci.* **27**, 890–895. (doi:10.1175/1520-0469(1970)027<0890:TMFTCO>2.0.CO;2)
6. Williamson DL. 2007 The evolution of dynamical cores for global atmospheric models. *J. Meteorol. Soc. Jpn B* **85**, 241–269. (doi:10.2151/jmsj.85B.241)
7. Cooley JW, Tukey JW. 1965 An algorithm for the machine calculation of complex Fourier series. *Math. Comput.* **19**, 297–301. (doi:10.1090/S0025-5718-1965-0178586-1)
8. Temperton C. 1983 Self-sorting mixed-radix fast Fourier transforms. *J. Comput. Phys.* **52**, 1–23. (doi:10.1016/0021-9991(83)90013-X)
9. Wedi NP, Hamrud M, Mozdzynski G. 2013 A fast spherical harmonics transform for global NWP and climate models. *Mon. Weather Rev.* **141**, 3450–3461. (doi:10.1175/MWR-D-13-00016.1)
10. Tygert M. 2008 Fast algorithms for spherical harmonic expansions, II. *J. Comput. Phys.* **227**, 4260–4279. (doi:10.1016/j.jcp.2007.12.019)
11. Tygert M. 2010 Fast algorithms for spherical harmonic expansions, III. *J. Comput. Phys.* **229**, 6181–6192. (doi:10.1016/j.jcp.2010.05.004)
12. O'Neil M, Woolfe F, Rokhlin V. 2010 An algorithm for the rapid evaluation of special function transforms. *Appl. Comput. Harmon. Anal.* **28**, 203–226. (doi:10.1016/j.acha.2009.08.005)
13. Cheng H, Gimbutas Z, Martinsson PG, Rokhlin V. 2005 On the compression of low rank matrices. *SIAM J. Sci. Comput.* **26**, 1389–1404. (doi:10.1137/030602678)
14. Coté J, Staniforth A. 1998 A two-time level semi-Lagrangian, semi-implicit scheme for spectral models. *Mon. Weather Rev.* **116**, 2003–2012. (doi:10.1175/1520-0493(1988)116<2003:ATTLSL>2.0.CO;2)
15. Hortal M. 1999 The development and testing of a new two-time-level semi-Lagrangin scheme (SETTLS) in the ECMWF forecast model. Technical report 292. European Centre for Medium-Range Weather Forecasts, Reading, UK.
16. Orszag SA. 1971 On the elimination of aliasing in finite-difference schemes by filtering high-wavenumber components. *J. Atmos. Sci.* **28**, 1074. (doi:10.1175/1520-0469(1971)028<1074:OTEOAI>2.0.CO;2)
17. Hortal M, Simmons AJ. 1991 Use of reduced Gaussian grids in spectral models. *Mon. Weather Rev.* **119**, 1057–1074. (doi:10.1175/1520-0493(1991)119<1057:UORGGI>2.0.CO;2)
18. Courtier P, Naughton M. 1994 A pole problem in the reduced Gaussian grid. *Q. J. R. Meteorol. Soc.* **120**, 1389–1407. (doi:10.1002/qj.49712051913)
19. Martinsson PG, Rokhlin V. 2007 An accelerated kernel-independent fast multipole method in one dimension. *SIAM J. Sci. Comput.* **29**, 1160–1178. (doi:10.1137/060662253)
20. Abdalla S, Isaksen L, Janssen PAEM, Wedi NP. 2013 Effective spectral resolution of IFS. *ECMWF Newsl.* **137**, 19–22.
21. Skamarock WC. 2004 Evaluating mesoscale NWP models using kinetic energy spectra. *Mon. Weather Rev.* **132**, 3019–3032. (doi:10.1175/MWR2830.1)
22. Lander J, Hoskins BJ. 1997 Believable scales and parameterizations in a spectral transform model. *Mon. Weather Rev.* **125**, 292–303. (doi:10.1175/1520-0493(1997)125<0292:BSAPIA>2.0.CO;2)