The ECMWF model: progress and challenges

Nils P. Wedi, Mats Hamrud, George Mozdzynski

ECMWF, Shinfield Park, Reading
RG2 9AX, United Kingdom
wedi@ecmwf.int

ABSTRACT

This paper reviews the current status of the ECMWF Integrated Forecasting System (IFS) model. In view of the challenges to run efficiently on the next generation of future computing architectures, the spectral transform technique and recent developments using the spectral computations, such as the fast Legendre transform (FLT), are reviewed. Early scientific results, obtained using the FLTs at ultra-high global resolution T_l3999, are very encouraging. However, there is a concern that spectral-to-gridpoint transpositions will be (energy-)inefficient given existing road maps to exascale computing, and in the longer term alternative data-structures and algorithms must be found.

1 Introduction

The dynamical core is an important element of every atmosphere or ocean model. Historically, in the context of numerical weather prediction and climate research, the dynamical core is the (dry) part of the model without any diabatic forcings (i.e. "physics"). However, with resolutions ranging from $\Delta = \mathcal{O}(0.1 - 20km)$ many processes that were traditionally computed in the subgrid-scale parameterisation ("physics") (e.g. gravity wave drag, convection, moist processes, boundary layer turbulence), are partially or fully resolved. As a result, it becomes increasingly ambiguous that a parametrized (subgrid-scale) term should be computed at all. Subsequently, the implications for the dynamics-physics interface of the model are the need to extend the dynamical core and to add new physical parameterizations. The subject is further discussed in other contributions to this seminar series, in particular in view of the use of either the non-hydrostatic or the hydrostatic dynamical core, cf. Malardel (2013) and Smolarkiewicz et al. (2013). The governing equations used can have a substantial impact on the efficiency of the dynamical core and its parallelisation strategy.

While the computational efficiency remains one of the most pressing needs of numerical weather prediction (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 time-stepping (Robert et al., 1972) and semi-Lagrangian advection (Ritchie, 1988), see also Diamantakis (2013)) there has been a steady increase in resolution, by approximately doubling the global horizontal resolution every 8 years at 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 about one day per decade (Simmons and Hollingsworth, 2002).

The factors driving continued horizontal resolution increases are: 1) at current resolutions important processes determining the vertical redistribution of energy in the atmosphere are not resolved and global NWP has reached the threshold of permitting and resolving convection explicitly; 2) 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; 3) so far horizontal resolution increases have

improved the skill of NWP and climate predictions; 4) larger problems scale better on massively parallel platforms. However in the future, 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.

The Integrated Forecasting System (IFS) is based on the spectral transform, semi-Lagrangian, semi-implicit (compressible) hydrostatic model with an option to use non-hydrostatic dynamics. The prognostic equations on the sphere of the IFS (and the code-shared ARPEGE model of Météo-France) dynamical core were derived under the philosophy of extending the hydrostatic primitive equations to the fully compressible Euler equations within the same overall numerical framework (Ritchie et al., 1995; Laprise, 1992; Bubnová et al., 1995; Temperton et al., 2001; Bénard et al., 2005, 2010; Wedi et al., 2009; Yessad and Wedi, 2011). The IFS is already a highly parallel application using an MPI/OpenMP hybrid parallelisation scheme (Barros et al., 1995). On standard CPUs, the IFS has been run at very high resolution ($T_l3999L137$) across 200K+ cores, cf. Mozdzynski (2013). Yet with emerging novel computing architectures, there is a need to increase substantially the level of fine-grain parallelism and local vectorisation, to exploit the speed-ups potentially available with accelerator-type (e.g. GPGPU, MIC) architectures. The preparation of the IFS for the next generation of HPC architectures with very large numbers of processors has been identified as essential for the future success of ECMWF. The issue of scalability poses a challenge. For the model, scalability issues are expected to become highly relevant for resolutions beyond T_l2047 (or equivalently $\Delta = 10$ km) with the use of spectral transforms because of significant single-processor memory and global communication requirements. Given the high accuracy and efficiency of the spectral transform method at hydrostatic scales, it is possible that energy efficient algorithms of the future exploit a hybrid strategy, where spectral transforms could still play an important role, but where for example derivatives may be calculated locally (with a small stencil of points), or where a lower dimension semi-implicit, hydrostatic solution acts as a preconditioner for non-hydrostatic algorithms. This article documents the existing efficiency of the spectral transform method in IFS. In particular, recent developments at ECMWF are described that mitigate the concern about the disproportionally growing computational cost of the spectral computations with increasing resolution (Wedi et al., 2013).

In IFS, horizontal resolution is expressed by the cut-off spectral truncation number N of the spherical harmonics series expansion of the prognostic variables. To illustrate where the cost per timestep is spent, Fig. 1 compares two simulations with the same number of gridpoints but spectral truncations according to (a) T_q1364 and (b) T_l2047. The T_q1364 refers to a quadratic grid (Orszag, 1971; Eliasen et al., 1970; Hortal, 1999) simulation whereas the T_l2047 refers to a linear grid (Coté and Staniforth, 1998; Hortal, 1999). The T_l2047 simulation is more expensive in the spectral part of the computations by approximately 30 percent. Both simulations use a reduced grid where the number of points is reduced towards the poles, keeping the relative distances between points approximately constant (Courtier and Naughton, 1994). Notably, about 60 percent 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 to 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 (in Fig. 1 a coarser grid corresponding to T_l799 has been used). 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 linear Gaussian grid is attractive when the wall-clock time cost is determined purely by the total number of gridpoints used, since N waves are used in the spectral transform to a gridpoint space with $2N+1$ gridpoints along a Gaussian latitude, compared to the quadratic grid, where N waves are used

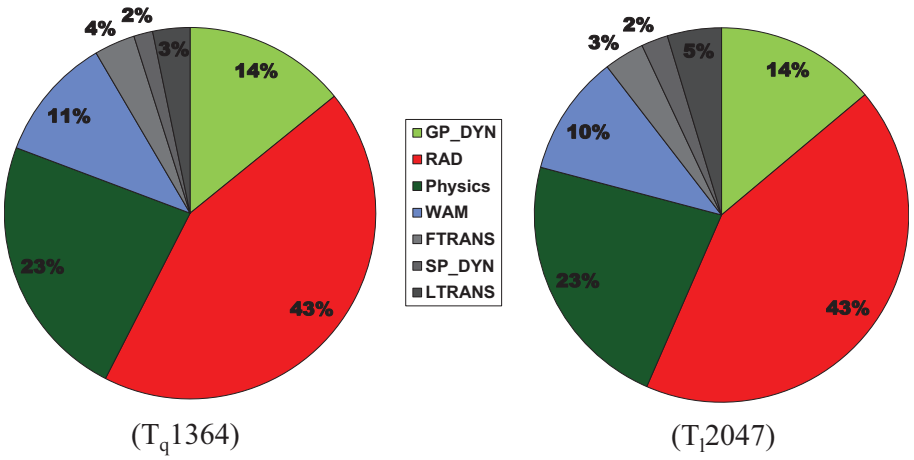


Figure 1: Cost distribution of a 10 day forecast at T_q1364 resolution (left) and the cost distribution at T_l2047 (right). Both forecasts use the same number of gridpoints. The dark colours ($\approx$ 10% of the total) represent computations associated with spectral space including the transpositions from gridpoint to spectral to gridpoint, namely 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 parameterisation 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 percent for the T_q1364 quadratic grid.

to transform to a gridpoint space with $3N+1$ gridpoints (Hortal, 1999). But on a grid with $2N$ points, Fourier components $exp(ikx)$ with $|k| > N$ will appear as low wavenumbers, the high wavenumbers are "aliased" to the low (Boyd, 2001, chap. 11). In the IFS model, aliasing will show up as "spectral blocking", a build up of energy at the smallest scales that may ultimately lead to instability. For example, quadratic terms (i.e. products of prognostic variables such as the pressure gradient term in the momentum equation) cannot be represented accurately beyond $2/3N$, because quadratic interactions will produce wave modes larger than N that alias onto the waves between $2/3N$ and N. A procedure that is devised to systematically eliminate quadratic aliasing makes use of the $2/3$ rule (Orszag, 1971). This rule is to filter all waves $|k| > (2/3)N$ since the quadratic interaction of two wavenumbers $|k| \leq (2/3)N$ will alias only to wavenumbers that will be purged by the filter. Notably, purging $1/3$ of the wave spectrum in spectral space is equivalent to filtering waves with wavelengths between 2Δ and 3Δ in a gridpoint model (Orszag, 1971). Using a quadratic grid with $3N+1$ gridpoints along a Gaussian latitude for N wave modes does this automatically, but with the disadvantage of increasing the number of points compared to the linear grid and thus the computational cost in gridpoint space. However, due to the increased cost of the Legendre transforms at resolutions with $N > 1000$, the situation is reversed and makes the quadratic or cubic grid an attractive alternative (Wedi, 2013).

The article is organised as follows: In the next section it is described how aliasing is controlled for the linear grid by horizontal diffusion and a special de-aliasing filter. Section 3 describes the spectral

transform method and the successful implementation of the Fast Legendre Transform (FLT) into the IFS. In section 4 an efficient spectral filtering technique is described that uses a Fast Multipole Method (FMM). This powerful algorithm has been implemented in IFS as part of the FLTs; it is no longer required for this application but may be useful for other applications in the future. Section 5 concludes the paper.

2 The control of aliasing

In IFS, a reduced grid is used (Hortal and Simmons, 1991), where the number of longitudes is reduced towards the poles, keeping the relative distances between points approximately constant, i.e. quasi-uniform. Towards the poles the number of gridpoints can be reduced owing to the property of the associated Legendre functions tending abruptly to zero towards the poles for large zonal wavenumber (see also the end of section 3). The reduction of the gridpoints and correspondingly the Fourier modes with increasing latitude away from the equator follows the rule with $3N_r + 1$ gridpoints along each reduced Gaussian latitude for the given number of waves N_r, see Courtier and Naughton (1994) for details. The reduction starts approximately outside of ± 30 degrees latitude. Thus all reduced latitudes eliminate aliasing in East-west direction from quadratic terms (even on the linear reduced grid) due to the particular choice of the reduced gridpoints and reduced Fourier modes. The $3N_r + 1$-rule is perhaps one of the reasons why the aliasing has not been so apparent when the linear grid was first introduced. Nevertheless, east-west aliasing of quadratic terms is still visible near the equator where the number of gridpoints is not reduced and matches the $2N + 1$ rule, i.e. $2N + 1$ gridpoints along equatorial latitudes to N wave modes, cf. panel a in Fig. 2. North-south aliasing can exist at all $2N + 1$ Gaussian latitudes with the linear grid, but is most visible in the vicinity of mountain ranges, e.g. the Alps, the Himalayas and the Andes, reflecting the influence of orography on the (quadratic) pressure gradient term. Panel a in Fig. 3 illustrates the north-south aliasing in the adiabatic meridional wind tendencies south of the Himalaya. The extend of the noise also spoils the adiabatic and physical tendencies archived as part of the special topic year-of-tropical-convection (YOTC) dataset where the physical parameterisation tendencies often compensate and thus equally show the aliasing noise with opposite sign. The aliasing worsens with increasing horizontal resolution since the absolute number of potentially aliased waves is growing (the last one third of the spectrum).

The term most responsible for the aliasing noise in IFS is the pressure gradient term in the momentum equation, although aliasing also exists due to the right-hand-side of the thermodynamic equation. However, here we shall focus on the aliasing from the momentum equation. Hence, in this de-aliasing procedure the difference between a filtered pressure gradient term and the unfiltered pressure gradient term is subtracted at every timestep from the right-hand-side of the momentum equation. The filtered term is obtained by computing the rotational and divergent components of the pressure gradient term in spectral space, smoothly truncate only the rotational component at approximately $2/3$ of the maximum truncation wavenumber N and transforming the result back to gridpoint space. This filter exploits the fact that aliasing projects entirely onto the rotational modes and hence only the rotational part of the pressure gradient term is filtered. The procedure may be understood in the spirit of ensuring mimetic properties with respect to the operation $\nabla \times \nabla \equiv 0$. The procedure eliminates the aliasing noise as evident in panels b of Fig. 2 and Fig. 3. Due to the compensation by the physics (and in particular the vertical diffusion scheme) aliasing has not been very visible so far in the diagnosis of the prognostic variables. However, Fig. 4 shows the spurious noise in the instantaneous vorticity field at 200 hPa which is eliminated by the de-aliasing procedure.

The truncation of the pressure gradient term has substantial influence on the kinetic energy spectra, which sharply decline beyond $2/3N$ when de-aliasing is applied. The aliasing and its removal can be clearly identified in Fig. 5. The energy spectra are strongly affected by the aliasing noise but have been dampened in the past by applying strong horizontal diffusion. In IFS, the diffusion of a variable ψ is

a b

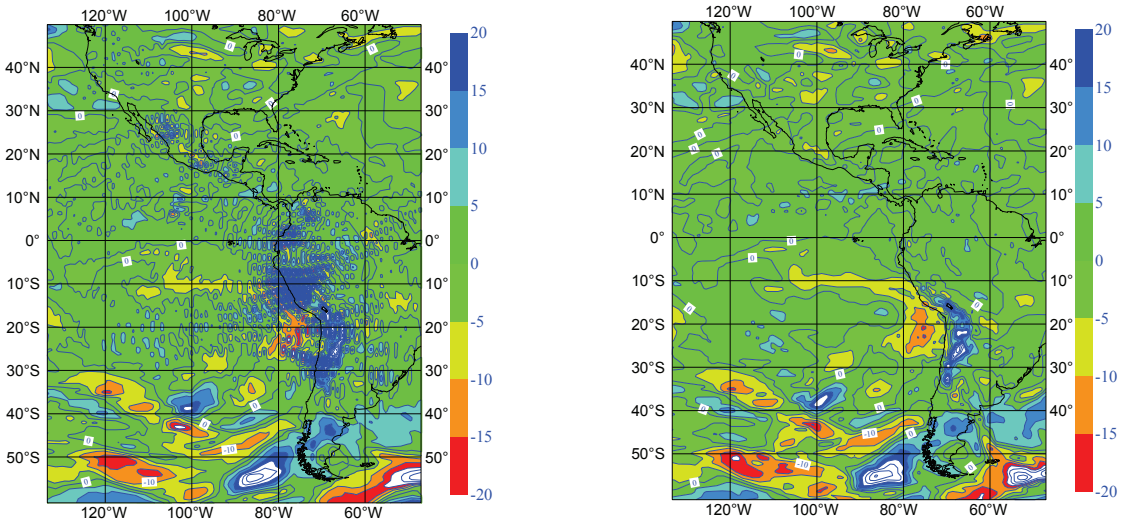
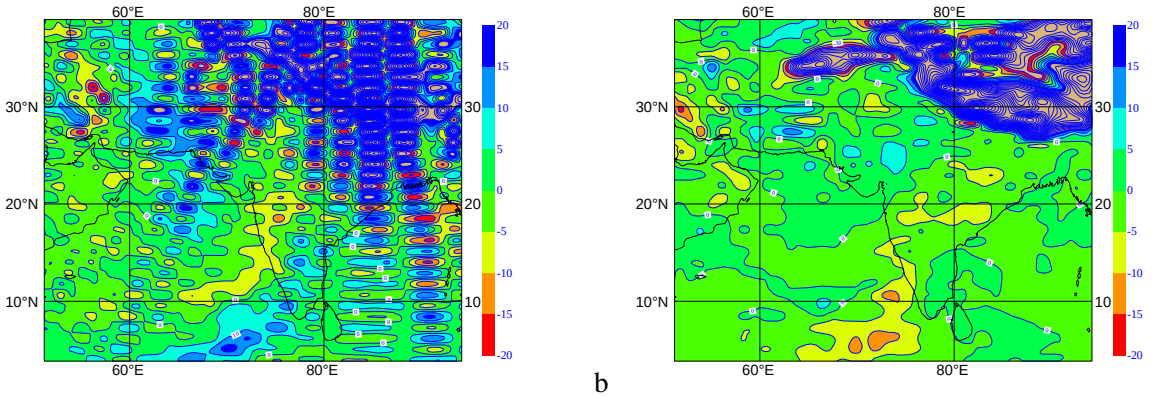
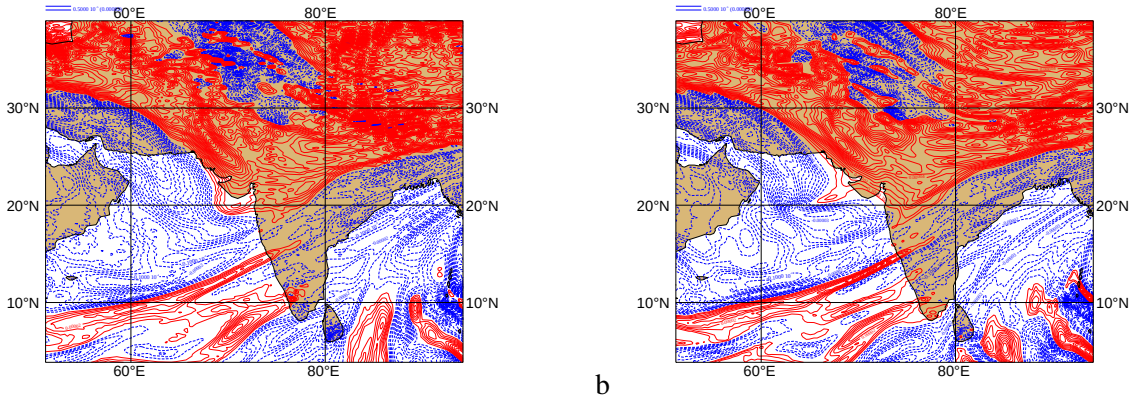


Figure 2: Panel a shows the east-west aliasing noise near the Andes in the 500hPa adiabatic zonal wind tendencies [m/s] used as input to the physical parametrizations. The zonal wind tendency is accumulated over 12 hours of the T159 simulation. Panel b shows the corresponding result when the de-aliasing procedure is applied at every timestep.



a b

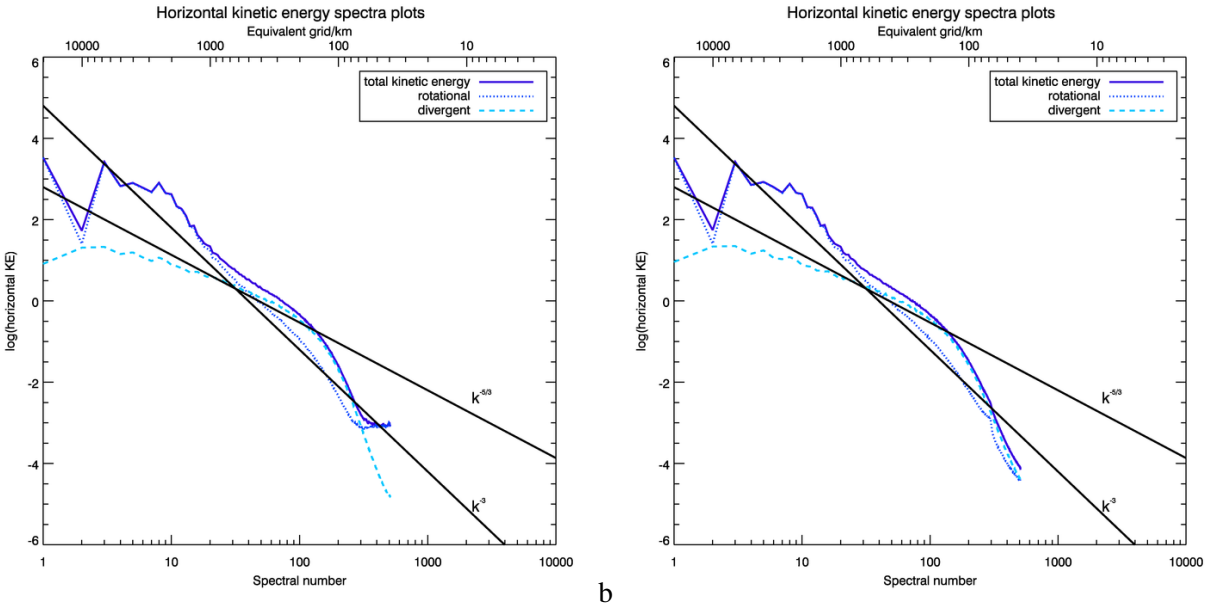
Figure 3: Panel a shows the north-south aliasing noise south of the Himalayas in the 700hPa adiabatic meridional wind tendencies [m/s] used as input to the physical parametrizations. The meridional wind tendency is accumulated over 24 hours of the T159 simulation. Panel b shows the corresponding result when the de-aliasing procedure is applied at every timestep.



a b

Figure 4: Panel a shows the aliasing noise seen in the vorticity field at 200hPa in a T511 simulation over the Himalayas region. Panel b shows the result after the de-aliasing procedure has been applied.

computed in spectral space by adding a term $-(-1)^r K\nabla^{2r}\psi$ with $r=2$, where the diffusion coefficient is given as $K=\tau^{-1}[a^2/(N(N+1))]^r$, noting that $\nabla^2\psi \equiv -[n(n+1)/a^2]\psi$. The time-scale of diffusion is denoted by τ. The de-aliased simulation in Fig. 5 used a longer time-scale for horizontal diffusion with $\tau = 6\Delta t$ (except in the sponge layer at the model top), compared to the aliased control with $\tau = 2\Delta t$. Given, that the IFS simulations are now essentially free of quadratic aliasing (higher order remains due to the thermodynamic equation), it has been found that the strong horizontal diffusion previously applied is no longer necessary and may even be detrimental to the medium-range forecast skill; except at the top of the model, where the spurious reflection of vertically propagating gravity waves is still effectively controlled by increased horizontal diffusion ("sponge" layer). The new de-aliasing filter has been found to be highly beneficial in removing spurious, near-surface oscillations and reducing the mass conservation error in the semi-Lagrangian advection by 50 percent.



a b

Figure 5: T511 kinetic energy spectra (solid) and its rotational (dashed) and divergent (dotted) components. Panel a shows the (aliased) control and panel b shows the de-aliased result. The spectra are shown at 100hPa and averaged over 10 days, discarding the first 12 hours.

The extra cost associated with the additional spectral transforms of the de-aliasing filter is approximately 5 percent of the overall model cost. Together with the non-linear model, tangent linear and adjoint model versions of the de-aliasing procedure have been developed. These are important since the latter models also suffer from aliasing noise, in part due to additional "quadratic terms" (products of background trajectory and perturbations) in the linearised equations.

3 Fast Legendre Transform

The spectral transform method has been successfully applied at ECMWF for the past thirty years, with the first spectral model introduced into operations at ECMWF 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. (1970) and Orszag (1970), 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 time-stepping scheme can be solved easily and horizontal gradients on the (reduced) Gaussian grid are computed accurately; particularly the Laplacian operator that is fundamental to the propagation of atmospheric waves. The success of the spectral transform method in NWP in comparison to alternative methods has been overwhelming, with many operational forecast centres having made the spectral transform their method of choice, as comprehensively reviewed in Williamson (2007).

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 (FFT) (Cooley and Tukey, 1965; Temperton, 1983) that reduces the computational complexity from $\mathcal{O}(N^3)$ to $\propto \mathcal{O}(N^2 \log N)$. However, with the conventional Legendre transform the spectral transform method is $\propto \mathcal{O}(N^3)$, 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. Due to the relative cost increase of the Legendre transforms compared to the gridpoint computations, very high resolution spectral models are believed to become prohibitively expensive, and methods based on finite elements or finite volumes on alternative quasi-uniform grids covering the sphere are actively pursued (e.g. Staniforth and Thuburn, 2012; Cotter and Shipton, 2012).

Legendre transforms involve sums of products between associated Legendre polynomials at given Gaussian latitudes and corresponding spectral coefficients of the particular field (such as temperature or vorticity at a given level). The FLT algorithm is based on the fundamental idea that for a given zonal wavenumber all the values of the associated Legendre polynomials at all the Gaussian latitudes of the model grid have similarities that may be exploited in such a way that one does not have to compute all the sums. Rather, FLTs precompute a compressed (approximate) representation of the matrices involved in the original sums and apply this compressed (reduced) representation instead of the full representation at every timestep of the model simulation. The FLTs are based on the seminal work of Tygert (2008, 2010), and in particular the algorithm for the rapid evaluation of special functions described in O'Neil et al. (2010) and the efficient interpolative decomposition (ID) matrix compression technique described in Cheng et al. (2005).

The computational cost of the conventional spectral transform method scales according to N^3 due to the cost of evaluating all the sums involved in the direct or inverse spectral transformation (Tygert, 2008). Since the fastest numerical methods used in geophysical fluid dynamics scale linearly with the number of gridpoints (i.e. proportional to N^2, recalling that the non-reduced linear grid has $(2N+1)\times(2N+1)/2$

gridpoints), the cost of the Legendre transforms would not be competitive at problem sizes $N = O(1000)$. Consequently, very high resolution spectral models may become prohibitively expensive. However, up to a resolution of approximately $N = 2047$ (10 km) 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. In all comparisons shown in this paper we have used the IBM cache/processor-optimised implementation of the matrix-matrix multiply routine *dgemm* from the IBM ESSL-library, which is substantially faster than a naive implementation of the required sums and which provides a stringent test for the possible speed-ups achieved by applying the FLT algorithm instead.

Considering that each field ζ represents l_{tot} vertical levels, we may write the inverse Legendre transform for each zonal wavenumber m as a matrix-matrix multiply problem of the form

$$\begin{pmatrix} \zeta_{m,1}(x_1) & \dots & \zeta_{m,l_{tot}}(x_1) \\ \vdots & \ddots & \vdots \\ \zeta_{m,1}(x_K) & \dots & \zeta_{m,l_{tot}}(x_K) \end{pmatrix} = \begin{pmatrix} \overline{P}_1^m(x_1) & \dots & \overline{P}_N^m(x_1) \\ \vdots & \ddots & \vdots \\ \overline{P}_1^m(x_K) & \dots & \overline{P}_N^m(x_K) \end{pmatrix} \begin{pmatrix} \zeta_{1,1}^m & \dots & \zeta_{1,l_{tot}}^m \\ \vdots & \ddots & \vdots \\ \zeta_{N,1}^m & \dots & \zeta_{N,l_{tot}}^m \end{pmatrix} \tag{1}$$

where the left-hand matrix represents a single Fourier coefficient ζ_m at all levels and at all Gaussian latitudes and the right-hand-side represents the matrix of the associated Legendre polynomial coefficients $\left[\overline{P}_n^m(x_k)\right]$ at given m for all total wavenumbers n and at all Gaussian latitudes, multiplied with the spherical harmonics coefficient matrix for a given m at all levels and for all total wavenumbers n. Notably, equation (1) exposes the parallelism of the problem, since all m, l are independent of each other and may be suitably distributed over the processors. A full description of the IFS parallelisation scheme is contained in Barros et al. (1995). Following Tygert (2010) the FLT algorithm is based on the fundamental idea that for a given zonal wavenumber m the matrix $\left[\overline{P}_n^m(x_k)\right]$ can be subdivided into smaller parts by doubling the columns and halving the rows with each level of subdivision and storing these level-by-level (of subdivision) in a hierarchical tree structure. This so called butterfly algorithm is described in detail in O'Neil et al. (2010). The essence of the algorithm is that each of the new sub-matrices obtained in this way may be compressed using an interpolative decomposition (ID) (Cheng et al., 2005) (or, in principle, any other algorithm for rank reduction), such that

$$|\mathscr{S}_{r\times s} - C_{r\times k}A_{k\times s}| \leq \varepsilon, \tag{2}$$

where matrix $C_{r\times k}$ constitutes a subset of the columns of sub-matrix $\mathscr{S}_{r\times s}$ and where matrix $A_{k\times s}$ contains a $k \times k$ identity matrix with k being called the ε-rank of sub-matrix $\mathscr{S}_{r\times s}$ (Cheng et al., 2005; Martinsson and Rokhlin, 2007). An important point to make is that the application of the ID compression directly on the full original matrices $\left[\overline{P}_n^m(x_k)\right]$ has no effect. Only when rank-deficient sub-matrices can be constructed by subdivision and by regrouping of columns and rows, the algorithm will save time and floating point operations. For NWP applications the advantage of the algorithm lies in the fact that one can precompute and store the compressed images of the sub-matrices. Moreover, in the application stage one merely needs to apply the compressed representation, resulting in far fewer summations that need to be executed at every time-step of the model simulation, provided that the rank-reduction is effective. The cost increase involved in the additional precomputation step is negligible, with typically less than 0.1 percent of the total elapsed time of a 10-day forecast simulation. Details of the implementation of the spectral transforms in IFS, the FLTs and the required pre-computations can be found in Wedi et al. (2013). In terms of computational cost we find up to resolutions of T7999 that the spectral transforms with FLTs scale according to $\mathscr{O}(N^2 \log^3 N)$.

FLTs represent a paradigm algorithm for trading formal accuracy and computational efficiency. The parameter ε defines the accuracy required in the compression part of the algorithm. In Wedi et al. (2013) the authors find that with $\varepsilon = 10^{-7}$ equivalent meteorological accuracy as measured in terms of hemispheric root-mean square error (rms) and anomaly correlation of 500 geopotential height and other parameters — 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 summarises these results. *Flop* refers to the counted number of floating point operations used in a 48h 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 7 independent selected dates. A stronger compression does reduce the computational cost further, but this does impact ultimately the meteorological forecast results negatively. 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 $\varepsilon = 10^{-10}$. Notably, we do not find any degradation in the meteorological result nor in global kinetic energy spectra when we reduce $\varepsilon = 10^{-4}$ for the inverse transform only (incidentally, this is the most costly part, since in this step also the derivatives are computed). In contrast, the direct transforms are sensitive to the choice $\varepsilon < 10^{-7}$. The instability reported in Wedi et al. (2013) with $\varepsilon = 10^{-2}$ can be eliminated if this threshold is only used for the inverse transform.

Table 1: Table of 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

The above results confirm the heuristic explanation why the butterfly algorithm works well for applications on the sphere. It is due to the structure of the spherical harmonics functions. Firstly, the amplitude of the spherical harmonics for large m is negligibly small towards the poles at many Gaussian latitudes — rows of the matrix $\left[\overline{P}_n^m(x_k)\right]$ to be subdivided — and the number of latitudes where the amplitudes are negligibly small is increasing with increasing truncation number N. Secondly, neighbouring n — represented by neighbouring columns of $\left[\overline{P}_n^m(x_k)\right]$ to be combined — are very similar. The former aspect is already used successfully in the application of the reduced grid mentioned earlier, where the number of points (and Fourier modes) are reduced at each latitude towards the poles, requiring ≈ 30 percent less gridpoints compared to an equivalent latitude-longitude grid and with the additional advantage of creating a quasi-uniform gridpoint distribution on the sphere, but without the penalty typically associated with reduced grids in other discretisations, e.g. finite-volume (Jablonowski et al., 2009).

4 Fast multipole methods and spectral filtering

The fast multipole method (FMM) was pioneered by Greengard and Rokhlin (1987) and provides a fast algorithm for the computation of a set of numbers f_j for all $j = 1, 2, ..., J$ defined as

$$f_j = \sum_{k=1}^{N} \frac{\beta_j P_k}{\tilde{\mu}_j - \mu_k} \tag{3}$$

with a specified precision of the computations at a cost proportional to $\mathcal{O}(J+N)$ instead of $\mathcal{O}(J \cdot N)$.

Using the Christoffel-Darboux formula for the associated Legendre functions, one can transform the problem of a direct and a subsequent inverse Legendre transform (in other words a spectral filter of a

physical field defined at a given set of latitudes) (Jakob-Chien and Alpert, 1997) as

$$\tilde{\zeta}^m(\tilde{\theta}_j) = \varepsilon^m_{N+1}\overline{P}^m_{N+1}(\mu_j)\sum_{i=1}^{J}\frac{\zeta^m(\theta_i)w_i\overline{P}^m_N(\mu_i)}{\tilde{\mu}_j-\mu_i} - \varepsilon^m_{N+1}\overline{P}^m_N(\tilde{\mu}_j)\sum_{i=1}^{J}\frac{\zeta^m(\theta_i)w_i\overline{P}^m_{N+1}(\mu_i)}{\tilde{\mu}_j-\mu_i} \tag{4}$$

The algorithm proposed in Yarvin and Rokhlin (1999) has been implemented in IFS but is not currently used, since the acceleration for the FLTs proposed in Tygert (2010) has been found to be very efficient and does not require the additional FMM algorithm (and its associated precomputations) originally proposed in Tygert (2008); at least for the spectral truncations used in NWP and climate simulations.

5 Conclusions

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. A dealiasing filter for the pressure gradient term in the momentum equation on the linear grid has been successfully implemented in IFS and works well at these resolutions. However, the quadratic grid may become an attractive alternative with the increasing cost of the Legendre transforms. Notably, the quadratic grid does not require any special dealiasing procedures.

The benefit of using the FLTs with enhanced compression in this resolution range is found to be limited, cf. Table 1. 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. However, the successful implementation of the FLTs mitigates the concern about the disproportionally growing computational cost of the Legendre transforms with increasing resolution. At T_l2047 ($\approx$ 10km), T_l3999 ($\approx$ 5km) and at T_l7999 ($\approx$ 2.5km) horizontal resolutions in actual NWP simulations, the spectral transform computations scale $\propto \mathcal{O}(N^2\log^3 N)$. For the simulations using the hydrostatic dynamical core, the MPI/OpenMP-parallelisation over the wavenumbers m still efficiently distributes the work, reducing the overall wall-clock time and the memory footprint on existing computing platforms. The forecast experience at globally ultra-high resolution T_l3999 and T_l7999 is naturally limited due to the high computational cost. However, there is some evidence, cf. (Wedi et al., 2012), that substantial benefits may be gained at these higher resolutions. One example are better predictions of significant wave height. Figure 6 shows the significant wave height, produced by the coupled wave model, at different horizontal resolutions during the landfall of hurricane Sandy, reaching New York on Monday night the 29th October 2012, where the T_l3999 simulation provides an excellent 3-day forecast. Improvements in associated realistic surface wind forcings and precipitation patterns (not shown) are also found.

The (energy-)cost and latency of the parallel communications associated with the gridpoint-to-spectral-to-gridpoint transforms and the communications (transpositions) within the spectral computations, have not been addressed in this article and remain a concern. Spectral-to-gridpoint transformations require data-rich (and energy-inefficient) global communications at every timestep that may become too expensive on future massively parallel computers. Techniques are currently explored to mitigate some of this cost, cf. Mozdzynski (2013).

In the longer term, however, in both the data assimilation and forecast model parts of the IFS system, improvements in scalability and energy efficiency will only be achieved with new scientific and algorith-

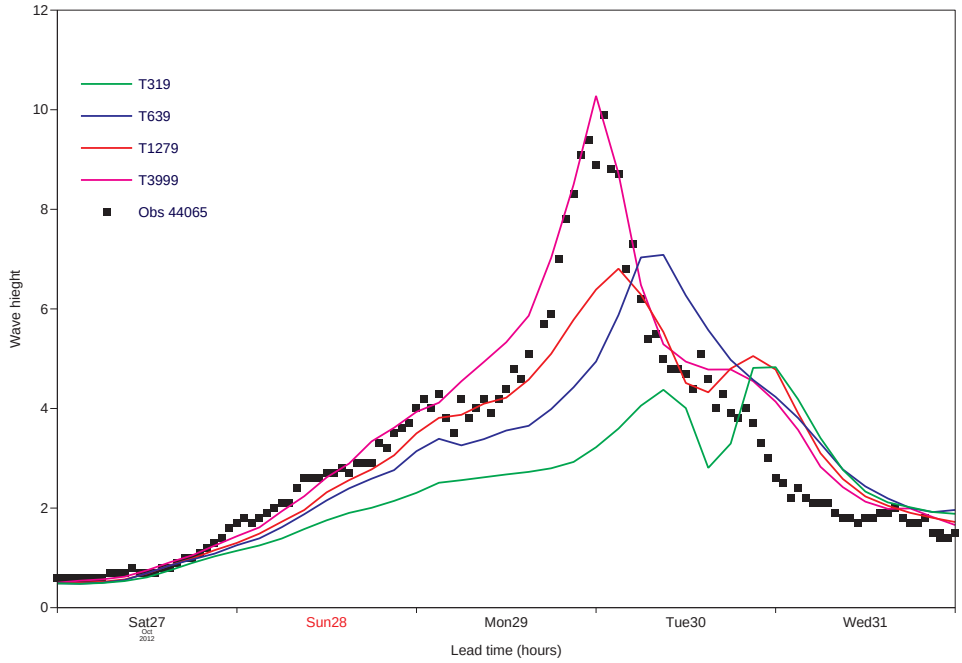


Figure 6: Significant wave height compared to a buoys observation during the landfall of hurricane Sandy near New York from a three-day forecast initialised on the 27th October 2012.

mic approaches and with alternative data structures using lower-level computational kernels, optimised for the emerging computer hardware technologies while minimising data movement.

Acknowledgements

We would like to thank Linus Magnusson and Jean Bidlot for providing an illustration of the potential of high resolution atmosphere simulations through forcing the ocean waves. Moreover, we would like to thank Christian Kühnlein and Piotr Smolarkiewicz for their comments on an earlier version of the manuscript.

References

Barros, S. R. M., D. Dent, L. Isaksen, G. Robinson, G. Mozdzynski, and F. Wollenweber (1995). The IFS model: a parallel production weather code. *Parallel Computing 21*, 1621–1638.

Bénard, P., J. Mašek, and P. Smolíková (2005). Stability of Leapfrog constant-coefficients semi-implicit schemes for the fully elastic system of Euler equations: case with orography. *Mon. Weather Rev. 133*, 1065–1075.

Bénard, P., J. Vivoda, J. Mašek, P. Smolíková, K. Yessad, C. Smith, R. Brožková, and J.-F. Geleyn (2010). Dynamical kernel of the Aladin-NH spectral limited-area model: Revised formulation and sensitivity experiments. *Q.J.R. Meteorol. Soc. 136*, 155–169.

Boyd, J. P. (2001). *Chebyshev and Fourier Spectral Methods* (2nd ed.). New York: Dover Publications Inc.

Bubnová, R., G. Hello, P. Bénard, and J.-F. Geleyn (1995). Integration of the fully elastic equations cast in the hydrostatic pressure terrain-following coordinate in the framework of the ARPEGE/Aladin NWP system. *Mon. Weather Rev. 123*, 515–535.

Cheng, H., Z. Gimbutas, P. Martinsson, and V. Rokhlin (2005). On the compression of low rank matrices. *SIAM J. Sci. Comput. 26*(4), 1389–1404.

Cooley, J. W. and J. W. Tukey (1965). An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation 19*, 297–301.

Coté, J. and A. Staniforth (1998). A two-time level semi-Lagrangian, semi-implicit scheme for spectral models. *Mon. Weather Rev. 116*, 2003–2012.

Cotter, C. J. and J. Shipton (2012). Mixed finite elements for numerical weather prediction. *J. Comput. Phys. 231*(21), 7076–7091.

Courtier, P. and M. Naughton (1994). A pole problem in the reduced Gaussian grid. *Q.J.R. Meteorol. Soc. 120*, 1389–1407.

Diamantakis, M. (2013). Semi-Lagrangian techniques for atmospheric modelling: current state and future challenges. Proc. ECMWF Workshop on Recent Developments in numerical methods for atmosphere and ocean modelling, Reading, UK. Eur. Cent. For Medium-Range Weather Forecasts. ibid.

Eliasen, E., B. Machenhauer, and E. Rasmussen (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.

Greengard, L. and V. Rokhlin (1987). A fast algorithm for particle simulations. *J. Comput. Phys. 73*, 325–348.

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, Eur. Cent. For Medium-Range Weather Forecasts, Reading, UK.

Hortal, M. and A. Simmons (1991). Use of reduced Gaussian grids in spectral models. *Mon. Weather Rev. 119*, 1057–1074.

Jablonowski, C., R. C. Oehmke, and Q. F. Stout (2009). Block-structured adaptive meshes and reduced grids for atmospheric general circulation models. *Phil. Trans. R. Soc. A 367*, 4497–4522.

Jakob-Chien, R. and B. Alpert (1997). A fast spherical filter with uniform resolution. *J. Comput. Phys. 136*, 580–584.

Laprise, R. (1992). The Euler equations of motion with hydrostatic pressure as an independent variable. *Mon. Weather Rev. 120*, 197–207.

Malardel, S. (2013). Physics/dynamics coupling at very high resolution. Proc. ECMWF Workshop on Recent Developments in numerical methods for atmosphere and ocean modelling, Reading, UK. Eur. Cent. For Medium-Range Weather Forecasts. ibid.

Martinsson, P. G. and V. Rokhlin (2007). An accelerated kernel-independent fast multipole method in one dimension. *SIAM J. Sci. Comput. 29*, 1160–1178.

Mozdzynski, G. (2013). Parallelisation and exascale computing challenges. Proc. ECMWF Workshop on Recent Developments in numerical methods for atmosphere and ocean modelling, Reading, UK. Eur. Cent. For Medium-Range Weather Forecasts. ibid.

O'Neil, M., F. Woolfe, and V. Rokhlin (2010). An algorithm for the rapid evaluation of special function transforms. *Appl. Comput. Harmon. Anal. 28*(2), 203–226.

Orszag, S. A. (1970). Transform method for calculation of vector coupled sums: application to the spectral form of the vorticity equation. *J. Atmos. Sci. 27*, 890–895.

Orszag, S. A. (1971). On the elimination of aliasing in finite-difference schemes by filtering high-wavenumber components. *J. Atmos. Sci. 28*, 1074.

Ritchie, H. (1988). Application of the semi-Lagrangian method to a spectral model of the shallow water equations. *Mon. Weather Rev. 116*, 1587–1598.

Ritchie, H., C. Temperton, A. Simmons, M. Hortal, T. Davies, D. Dent, and M. Hamrud (1995). Implementation of the semi-Lagrangian method in a high resolution version of the ECMWF forecast model. *Mon. Weather Rev. 123*, 489–514.

Robert, A., J. Henderson, and C. Turnbull (1972). An implicit time integration scheme for baroclinic models of the atmosphere. *Mon. Weather Rev. 100*, 329–335.

Simmons, A. J. and A. Hollingsworth (2002). Some aspects of the improvement in skill of numerical weather prediction. *Q.J.R. Meteorol. Soc. 128*, 647–677.

Smolarkiewicz, P. K., C. Kühnlein, and N. P. Wedi (2013). A unified framework for integrating sound-proof and compressible equations of all-scale atmospheric dynamics. Proc. ECMWF Workshop on Recent Developments in numerical methods for atmosphere and ocean modelling, Reading, UK. Eur. Cent. For Medium-Range Weather Forecasts. ibid.

Staniforth, A. and J. Thuburn (2012). Horizontal grids for global weather and climate prediction models: a review. *Q.J.R. Meteorol. Soc. 138*, 126.

Temperton, C. (1983). Self-sorting mixed-radix fast Fourier transforms. *J. Comput. Phys. 52*, 1–23.

Temperton, C., M. Hortal, and A. Simmons (2001). A two-time-level semi-Lagrangian global spectral model. *Q.J.R. Meteorol. Soc. 127*, 111–127.

Tygert, M. (2008). Fast algorithms for spherical harmonic expansions, II. *J. Comput. Phys. 227*, 4260–4279.

Tygert, M. (2010). Fast algorithms for spherical harmonic expansions, III. *J. Comput. Phys. 229*, 6181–6192.

Wedi (2013). Increasing horizontal resolution in NWP and climate simulations — illusion or panacea ? *Phil. Trans. R. Soc. A*. submitted.

Wedi, N. P., M. Hamrud, and G. Mozdzynski (2013). A fast spherical harmonics transform for global NWP and climate models. *Mon. Weather Rev. 141*, 3450–3461.

Wedi, N. P., M. Hamrud, G. Mozdzynski, G. Austad, S. Curic, and J. Bidlot (2012). Global, non-hydrostatic, convection-permitting, medium-range forecasts: progress and challenges. *ECMWF Newsletter 133*, 17–22.

Wedi, N. P., K. Yessad, and A. Untch (2009). The nonhydrostatic global IFS/ARPEGE: model formulation and testing. Technical Report 594, Eur. Cent. For Medium-Range Weather Forecasts, Reading, UK.

Williamson, D. L. (2007). The evolution of dynamical cores for global atmospheric models. *J. Meteor. Soc. Japan 85B*, 241–269.

Yarvin, N. and V. Rokhlin (1999). An improved fast multipole algorithm for potential fields on the line. *SIAM J. Numer. Anal. 36*, 629–666.

Yessad, K. and N. P. Wedi (2011). The hydrostatic and non-hydrostatic global model IFS/ARPEGE: deep-layer model formulation and testing. Technical Report 657, Eur. Cent. For Medium-Range Weather Forecasts, Reading, UK.