TECHNICAL MEMORANDUM

760

The modelling infrastructure of the Integrated Forecasting System: Recent advances and future challenges

N.P. Wedi, P. Bauer, W. Deconinck, M. Diamantakis, M. Hamrud, C. Kühnlein, S. Malardel, K. Mogensen, G. Mozdzynski, P.K. Smolarkiewicz

Research Department

November 2015

Special topic paper presented at the 44th session of ECMWF's Scientific Advisory Committee, Reading,UK

European Centre for Medium-Range Weather Forecasts
Europäisches Zentrum für mittelfristige Wettervorhersage
Centre européen pour les prévisions météorologiques à moyen

Series: ECMWF Technical Memoranda

A full list of ECMWF Publications can be found on our web site under:
http://www.ecmwf.int/en/research/publications

Contact: library@ecmwf.int

European Centre for Medium Range Weather Forecasts
Shinfield Park, Reading, Berkshire RG2 9AX, England

Abstract

With the objective to develop and maintain one of the most advanced and flexible modelling infrastructures in Europe for operational, global NWP applications, recent advances and future challenges are described. A particular challenge arises from the need to achieve computationally and energy efficient solutions for operating global, complex, high-resolution, ensemble based systems on high-performance computers so that they will remain affordable given tight operational schedules. In particular, the rising cost of quantifying uncertainty needs to be addressed. This paper presents the current status and steps taken towards increasing the model realism and complexity to improve forecasts with a sustainable modelling infrastructure. The progress to date includes the flexibility to explore unstructured horizontal discretizations, the addition of a new, powerful 3D solver for elliptic problems arising from the implicit discretization of the non-hydrostatic system, an option for inherently conserving, monotone, multi-tracer transport, and developments towards a flexible vertical coordinate formulation. Future scientific priorities are to combine the strengths of the newly developed, principally autonomous non-hydrostatic finite-volume module FVM with the hydrostatic semi-Lagrangian spectral transform options of the IFS, to review the vertical discretization, and to carefully address physics-dynamics as well as Earth-system component coupling.

1 Introduction

The Integrated Forecasting System (IFS) comprises a substantial set of routines to perform data assimilation and to run forecasts ranging up to one year ahead, as well as routines for observational data and model output handling. While the number of actively assimilated observations is expected to grow by one order of magnitude in the next decade, global model resolution and complexity are expected to increase by at least two orders of magnitude. Enhancing model complexity where it improves forecast skill includes adding actively interacting chemical processes in the atmosphere and fully coupled simulations of the atmosphere with ocean, sea-ice and land surfaces. All these components share routines from what is called here the model infrastructure of the IFS. A distinct component of the model infrastructure is the so called dynamical core that treats the interaction of the resolved motions including all adiabatic processes. The dynamical core is the spine of any numerical weather prediction (NWP) system and must satisfy stringent requirements regarding its accuracy, stability, and efficiency in terms of time-to-solution. In particular, given suitable initial and boundary conditions combined with sufficient resolution, a (moist) dynamical core alone could provide a meaningful prognosis of global weather evolution. However, 'sufficient resolution' for practical NWP purposes may be in the 10-100m range and thus proves unaffordable. Models employ physical parameterizations to describe the effect of unresolved processes on the resolved scale, but also to describe diabatic effects such as radiation and water phase changes.

As the limit between resolved motions and the sub-gridscale shifts with increasing horizontal resolution, the relative importance of the dynamical core will amplify (the stratospheric dynamics are already a good example) together with the need for more scale-aware parameterizations to represent partly resolved processes. As a result, the dependence on uncertain parameterizations is reduced, while the realism in areas with orography improves model variability, e.g. for ensemble systems.

The global kinetic energy distribution in the atmosphere depends on both wave propagation and material motion, which are handled by the dynamical core. Material motion forms a substantial contribution to the non-linear spectral transfer fluxes of energy between a given wavenumber and its neighbouring wave numbers (Augier and Lindborg, 2013; Malardel and Wedi, 2015). Figure 1 illustrates the direction of these non-linear spectral transfer fluxes (or in other words how different scales of motion influence each other) with values above zero indicating a downscale influence and values below the zero line indicating an upscale energy cascade. With increasing resolution more small-scale processes are resolved and it is important to monitor their upscale or downscale influence and the associated error growth. Notably, the total conversion of available potential energy (APE) to kinetic energy (KE) in the global atmosphere is only 0.26% of the incoming solar radiation at the top of the atmosphere (green line in Figure 1), which explains why numerical procedures employed in dynamical cores are subject to intricate conditions for both stability and accuracy.

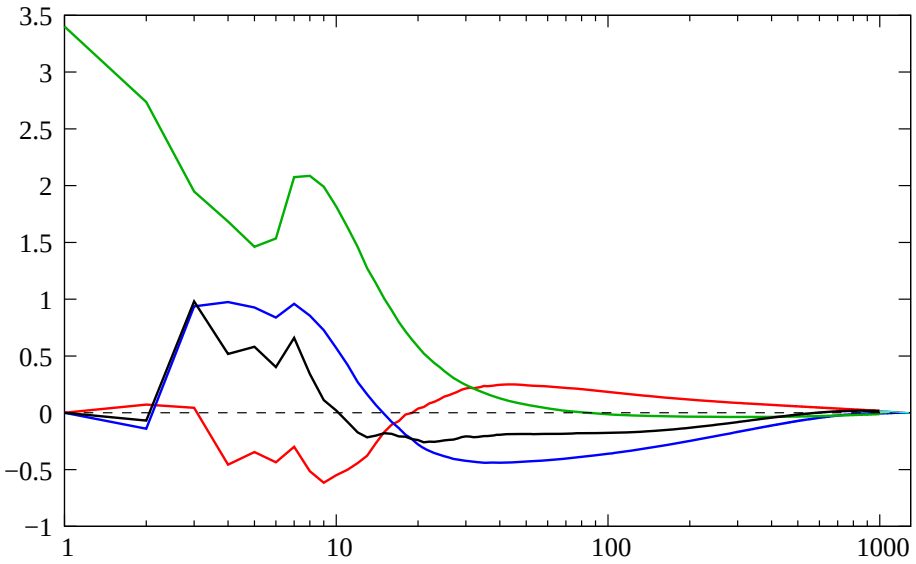

Figure 1: Globally integrated monthly average non-linear spectral fluxes in Watts/m^2 versus total wavenumber for January 2014 for the IFS ~9km horizontal resolution three day forecasts. Fluxes are accumulated from right to left with the red line denoting the kinetic energy (KE) flux, the blue line denoting the available potential energy (APE) flux and black is the sum. The green line denotes the accumulated total conversion of APE to KE which is only 0.26 percent of the incoming solar radiation, see Malardel and Wedi (2015) for more details.

The present factors driving continued horizontal resolution increases are: 1) at current resolutions important processes determining the vertical redistribution of energy in the atmosphere (e.g. convection, boundary layer turbulence) are not resolved but these are important for predicting high impact, extreme weather features; 2) more accurate resolved representations of the forcing, e.g. topography, vegetation, land-use, ocean currents and associated SSTs have a decisive impact on the atmospheric dynamics; 3) so far horizontal resolution increases have improved the skill of NWP and climate predictions (through advances in model and analysis accuracy together with better exploitation of dense observation networks and satellite observations) and from a computational perspective models with increased resolution still scale well. Future model development will also be driven by other factors: these are, from a scientific point of view, the reliability of forecasts together with a quantitative assessment of the uncertainty, and from a technical point of view, the energy efficiency and the (hardware-related) reliability of massively parallel computations (Wedi, 2014). In particular the technical aspects of massively parallel computations on emerging hardware may constrain the algorithmic choices or create additional challenges to explicitly handle the energy consumption or functionality of the compute unit for particular processes.

Substantial challenges lie ahead due to the rising cost of energy associated with running complex high-resolution forecast models with massively increasing concurrency required, combined with the likelihood that Moore's law will soon saturate or slow down, whereby microprocessor density (and performance) no longer doubles every 18 months. The biggest challenge for state-of-the-art computational NWP services arises from its own software productivity shortfall. The application

software at the heart of all services at the European Centre for Medium-Range Weather Forecasts (ECMWF) is inadequately equipped to efficiently adapt to the rapidly evolving heterogeneous hardware provided by the supercomputing industry. ECMWF is not alone in facing these challenges and has initiated a scalability programme of targeted activities in collaboration with national meteorological services, high-performance computing (HPC) centres, vendors and academia.

Considering the substantial effort required to develop a new model infrastructure, several other NWP centres have already started to explore alternatives. The motivation behind these developments are: efficiency on emerging energy-efficient high performance computing platforms, a better representation of conservation properties that are increasingly important for modelling the Earth system, an advanced coupling of the atmosphere and ocean, and a flexible regionalization through limited area nesting directly within the same discretization of the global domain.

The purpose of this paper is to give an overview of the current state-of-the-art IFS model infrastructure for global operational NWP at ECMWF, to describe recent progress, and to give an outlook in view of the future challenges.

In view of the above development priorities, section 2 describes the strengths and the need for development of the IFS in selected areas, namely computational affordability, intelligent data placement expressed through grid choices, and conservation properties. Recent advances to support the next horizontal resolution upgrade(s) are also described. In section 3, relevant developments and advances at other global NWP centres are summarized. Section 4 provides a roadmap of developments for the next three years. Section 5 details recent progress on several key development areas at ECMWF towards a flexible, scalable and sustainable model infrastructure. Finally section 6 concludes the paper.

2 The Integrated Forecasting System (IFS)

2.1 Assets

The spectral transform model

The spectral transform method has been successfully applied at ECMWF for approximately 30 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 parameterizations are computed. Unconditional stability and good dispersion properties of the semi-Lagrangian (SL) advection scheme in combination with unconditional stability of the semi-implicit (SI) time discretization (where the resulting Helmholtz problem is solved in spectral space) implies that the only limiting factor on the time step is the magnitude of local truncation errors.

The accurate space discretization scheme used in IFS combined with a fully-centred, second-order, SI time-stepping facilitates the use of very long time steps. The use of the spectral transform method in global NWP in comparison to alternative methods has been wide spread, with many operational forecast

centres having made the spectral transform approach their method of choice, as comprehensively reviewed in Williamson (2007). At ECMWF both the increased computing capacity and the corresponding advances in the numerical techniques applied --- namely the SI time stepping (Robert, 1972) and SL advection (Ritchie, 1988; Temperton et al., 2001) ---have led to a steady increase in horizontal resolution. At ECMWF the global horizontal resolution has approximately doubled every 8 years. This rate reflects the corresponding increase in necessary computing power and together with algorithmic and model advances has provided the basis for the improvement of medium-range predictive skill by 1 day per decade.

Due to the relative cost increase of the Legendre transforms compared to the gridpoint computations with increasing resolution, very high resolution spectral models were believed to become prohibitively expensive. However, the recent implementation of a fast Legendre transform (FLT) (Wedi et al., 2013) mitigates the concern about the disproportionally growing computational cost. Moreover, it has been found that the efficiency and accuracy of the hydrostatic, semi-implicit semi-Lagrangian (SISL) solution procedure using the spectral transform method can be enhanced substantially at higher resolutions by moving to the cubic truncation spherical harmonics grid (Wedi, 2014), where aliasing is avoided by oversampling with at least four gridpoints to one wavenumber, see section 2.2.2 below for more details on grid choices. The efficiency gain between the cubic transform and the linear transform grid can be estimated since the spectral transforms scale approximately with $M^2\,LOG^3\,M$, where M denotes the cut-off truncation wave number, and the cubic transform grid simply halves M compared to the corresponding linear grid.

Scalability of IFS with hybrid message passing using MPI/OpenMP

The current hybrid MPI/OpenMP implementation of IFS has been successfully used for about two decades. MPI (message passing interface) has been used primarily for the communication of data between tasks located across the distributed memory nodes of today's supercomputer systems. Details on the MPI implementation are given in Barros et al. (1995) and some aspects are emphasized in the subsection on the *Atlas* development below.

OpenMP on the other hand is used for distributing work across the processor cores within the shared memory space of each MPI task. The combination of MPI and OpenMP is important as this provides for good performance and memory utilisation today as shown in Figure 2 for the IFS simulation with truncation M=511 running on 48 CRAY XC-30 nodes, where 8 OpenMP threads per task delivers the optimal performance for this resolution. For ECMWF's XC-30 cluster systems, each compute node has about 54 GB available for use by IFS applications.

At resolutions above truncation M=511 it is no longer possible to run an IFS model using MPI-only (i.e. no OpenMP, or the leftmost point in Figure 2) without using fewer processor cores than the 24 cores that are available on an XC-30 node. This is mainly due to a combination of the IFS application code and the size of the executables for which there is a copy per MPI task.

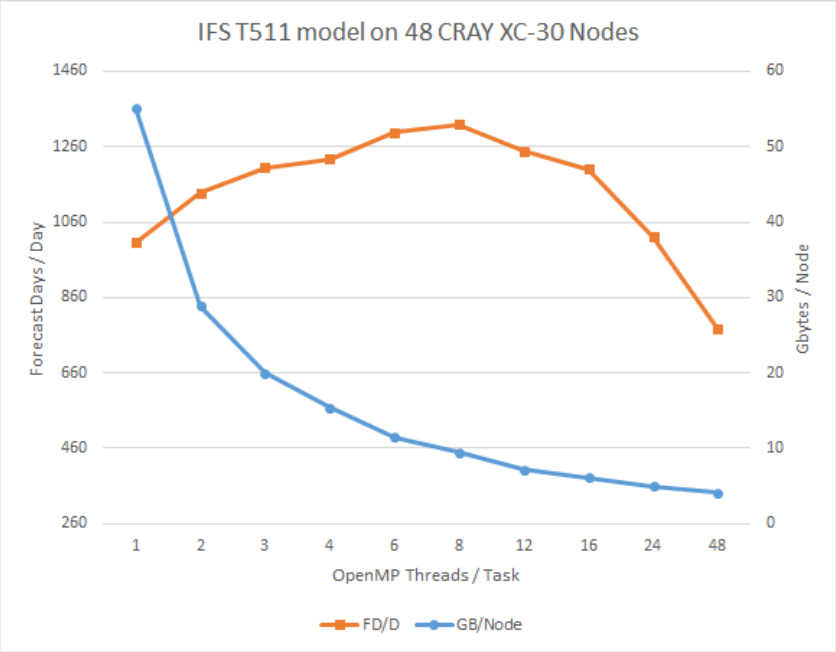

Figure 2: Performance and memory utilisation of an IFS TL511 model on 48 CRAY XC-30 nodes. The blue curve denotes Gigabytes of required memory per node whereas the red curve shows simulated forecast days per wall clock time day.

An example of a code area that does not memory-scale is the semi-Lagrangian transport scheme (see also sections 5.4 and 5.5) where relatively large halos are allocated which are derived from a maximum expected wind speed and the model time-step. The resulting halo area is therefore of constant size.

So what limits OpenMP thread scalability? There are a number of reasons for this limitation, including implicit data communication (moving data between different OpenMP parallel regions), the lack of sufficient work per thread, memory bandwidth limitations, and load imbalance to name a few. Nevertheless, the number of OpenMP threads per task has increased to 6 or 8, and we expect this trend to continue in the future, as processor sockets and accelerators contain increasing numbers of cores and cache.

Today's IFS model can scale well up to a limit of around 100 grid columns per thread. So for the TL1279 model there is an upper limit of efficiency at approximately 450 Cray XC-30 nodes (10800 cores), whereas we use 120 Cray XC-30 nodes for the model in operational production. For the next planned upgrade to a TCo1279 model, there is an upper limit of about 1,375 XC-30 nodes (33000 cores), whereas we currently expect to require approximately 400 Cray XC-30 nodes to fulfil the time-critical schedule in operations in early 2016.

To date we have only seen IFS scale to an upper limit of around 200 thousand cores on the ORNL TITAN machine used in the EU funded project (2011-2014) on Collaborative Research into Exascale Systemware, Tools & Applications (CRESTA, http://www.cresta-project.eu/). To get beyond this level of scalability will require the efficient use of new accelerator technologies, improved hierarchical memory management, further minimizing communication, and increased (low/thread)-level parallelism, together with changes in the mathematical algorithms used. However, these are complementary to improved standards of MPI and OpenMP that will continue to be used widely on future systems.

A major reason for continued use of MPI and OpenMP is that they are standards, and as such IFS can remain portable to different vendor systems and in particular systems that are used by our IFS partner Météo-France and other ECMWF member states. Moreover, there are plans to improve the resilience of these parallelisation and communication standards which will in return substantially enhance the failure resilience of NWP applications.

Within the CRESTA project ECMWF has explored overlapping computations and communications using Fortran2008 coarrays in the context of OpenMP parallel regions (Mozdzynski et al., 2015). In the future we expect that this approach can be used more generally (i.e. more portable), i.e. by providing a GASPI/GPI library interface (developed by the Fraunhofer Institute ITWM for portable error-resilient communication) as an alternative option to using Fortran2008 coarrays. The GASPI/GPI interface is further expected to become part of a future MPI standard.

2.2 The need for development

Computational affordability

The progress made in developing the spectral transform technique ensures its viability for operational applications for at least another decade. However, this depends also on the ability of emerging computing architectures to provide energy-efficient inter-node (MPI) communication. This is in particular true for the high memory requirements, the data-rich communication and the latency of the parallel communication within the spectral transforms. The progress in overlapping communication and computation may further the efficiency of SISL models, but ultimately the communication overhead and not the computational burden may be limiting the applicability of the spectral transform method. This is illustrated in Figure 3 for the 5km and 2.5km horizontal resolution IFS simulations, where the communication cost amounts to 75% of the cost within the transform part of the model on TITAN (no 2, top 500 list, June 2015). On the Cray XC-30 this is substantially improved but still 55 percent (not shown).

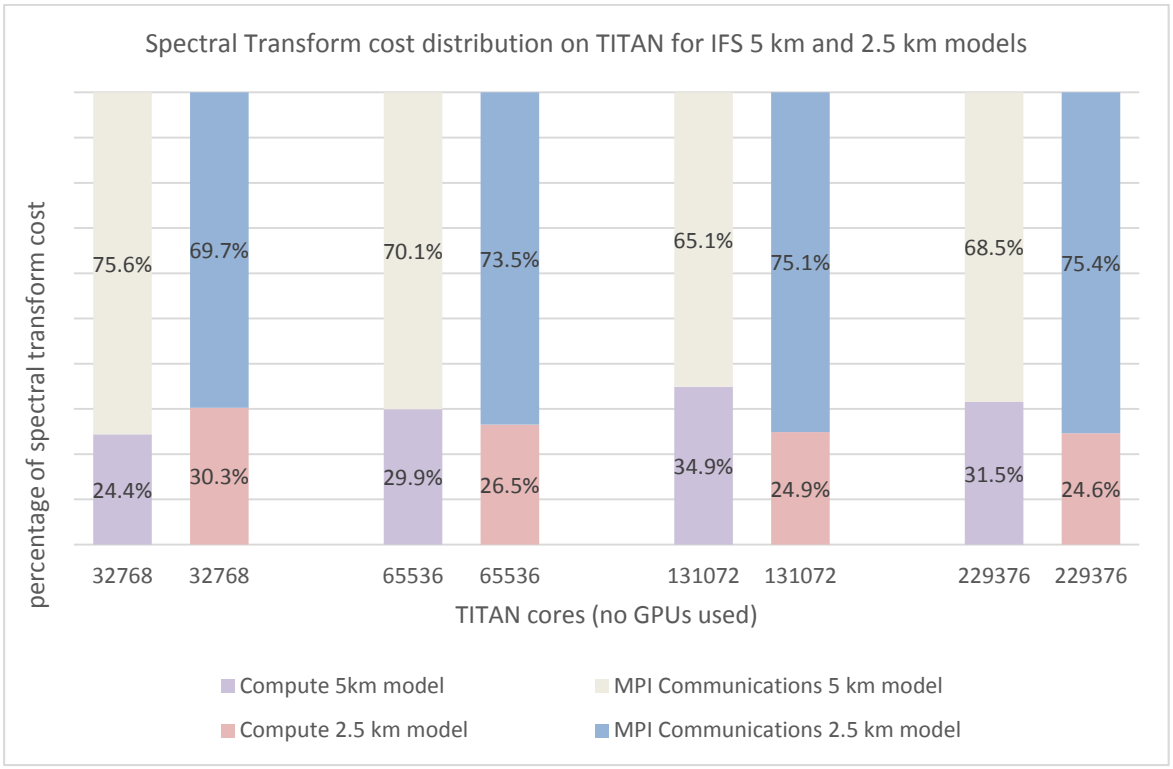

Figure 3: Spectral transform cost distribution on TITAN for IFS 5 km and 2.5 km models. The total cost of the spectral transforms on TITAN as a percentage of the overall wall clock time of the model simulation was in the range 29% to 51%.

The computational affordability of global non-hydrostatic forecasts is not merely determined by (hardware) technological advances. Replacing the highly efficient hydrostatic solution procedure (H-IFS) that has been central to the success of global weather and climate prediction comes at a price. Although IFS has a non-hydrostatic option (NH-IFS, Bubnová et al., 1995; Bénard et al., 2010; Wedi et al., 2009; Wedi and Smolarkiewicz, 2009; Yessad and Wedi, 2011) and is successfully used operationally at Météo-France (AROME) and several Aladin/HARMONIE partners for their limited-area forecasts with resolutions O(1km), the current cost in the global context is at least twice the cost of the global hydrostatic model, and there are potential limitations in the current formulation of the equations and the choice of discretization. These are further discussed in sections 5.4 and 5.6.

Grid choices

In spectral transform models resolution may be defined by the truncation wave number M or according to how many gridpoints represent the smallest representable wavelength in spectral space. Since some processes are entirely described in gridpoint space only (e.g. the surface scheme and non-linear advection) and some on different grids (e.g. radiation, wave model), there is not a single number that describes the model's resolution. Considering only regular (non-reduced) grids for the moment, a linear grid samples the smallest wavelength $2\pi a/M$ (where a is the Earth's radius) by 2 points, a quadratic grid by 3 points and a cubic grid by 4 points. The choice is dictated by the need to accurately represent (i.e. "alias free") linear, linear and quadratic, or linear, quadratic and cubic product terms in the

equations. Recent experience suggests that the importance of non-linear forcing terms on the right-hand side of the equations increases with increasing resolution, thus exacerbating the problem of aliasing and requiring the need to review the grid choice. So far either a quadratic or a linear grid has been used at ECMWF, but experiments at higher resolution suggest distinct advantages, both for efficiency and accuracy, of a cubic grid (Wedi, 2014). In particular, aliasing effects are avoided, global mass conservation is superior (cf section 2.2.3), and the effective resolution is improved as measured by improved predictive skill together with a more realistic kinetic energy spectrum (cf section 5.7.1).

In practice, a reduced grid is used (Hortal and Simmons, 1991; Courtier and Naughton, 1994), where the number of longitudes is reduced towards the poles, keeping the relative distances between points approximately constant, i.e. quasi-uniform. This reduces the number of gridpoints by approximately 30-35%. However, additional efficiency savings can be made by introducing a new (cubic) octahedral reduced Gaussian grid (cf section 5.2), where the number of points is further reduced by 22%. The latter relaxes the uniformity of grid-distances but instead introduces a smooth transition of nearly uniform mesh elements, constructed by triangular vertices connecting the Gaussian gridpoints and then creating the corresponding median-dual mesh (Figure 11). The median-dual mesh is an alternative mesh arrangement distinct from the Voronoi approach, and provides flexibility for arbitrarily shaped tessellations (Szmelter and Smolarkiewicz, 2010). This further enhances the efficiency without compromising the global solution accuracy or the ability to perform spectral transforms, and provides a route for accurate local derivative computations on the median-dual meshes associated with the Gaussian gridpoints. Moreover, it provides a pathway for exploring hybrid methods, where progressively more aspects of the solution are solved locally in gridpoint space, thus reducing communication requirements, cf sections 5.1 and 5.4.

For convenience we introduce the following nomenclature for truncations and grids. TLXX denotes a linear (reduced) Gaussian grid with triangular truncation M=XX. TCXX denotes a cubic (reduced) Gaussian grid with triangular truncation M=XX. TCoXX denotes a cubic octahedral mesh with triangular truncation M=XX. NYY denotes a Gaussian grid with the number of latitudes YY between pole and equator.

Mass conservation

The lack of inherent mass conservation is a weakness in IFS and a typical problem in standard SL schemes employed by operational NWP models. Mass conservation problems associated with SL schemes are suspected to be the reason for some biases in the upper troposphere lower stratosphere (UTLS) region. Fully consistent, inherently mass conserving SL schemes exist, e.g. SLICE (Zerroukat and Allen, 2012) and CSLAM (Lauritzen et al., 2010). However, mass conservation is achieved at a high computational expense which tends to be prohibitive for operational applications, given that there is only a small increase expected in the skill of the medium-range forecast. Furthermore, as inherently mass conserving SL methods need to trace grid cell volumes back in time using the long timesteps of global NWP models, their accuracy in areas of strong flow deformation (e.g. near complex terrain) deteriorates and spurious divergence is generated that may lead to worse results than the lack of formal mass conservation. The main problem here is that SL trajectories are computed using a purely kinematic algorithm that is unconstrained by the continuity of the prevailing flow (Cossette et al., 2014). A step towards improving local conservation in this spirit has been recently taken with the introduction of the

COMAD interpolation algorithm in the IFS (Malardel and Ricard, 2015), where a modification of the interpolation weights in the SL transport scheme introduces the concept of cell-averaging into the traditional pointwise SL scheme, thus improving the continuity and the conservative property of the re-mapping between the model gridpoints and the origin points of the backward trajectories.

Although alternatives to the existing SL scheme exist, such as the locally mass conserving semi-Lagrangian (LMCSL) scheme (Kaas, 2008), the implementation in the IFS would require a general change of variables. This implies a substantial code modification in the current dynamical core, its interface with the parametrisations, and its interaction with the SI scheme, cf. Malardel and Ricard (2015) for a discussion. The problem exposes the inflexibility and the necessary planning effort involved when attempting seemingly small changes with a highly tuned monolithic code structure, be it for scientific research or for operational applications beyond the standard forecast and assimilation purpose. Hence, the required flexibility and availability of alternative transport options within IFS is highly desirable, also in view of atmospheric composition forecasts as part of the Copernicus services, and further discussed in sections 5.4 and 5.5. Given the modern approaches and high level languages that are employed in reorganising legacy codes such as IFS, the additional flexibility should no longer mean a loss of efficiency.

Notably, even though the SL advection scheme in IFS is not inherently mass conserving, the mass conservation error for the total mass of air is small. The current level of mass conservation error for the total mass of air is 0.005% of initial mass for a 10-day forecast with the next operational octahedral Gaussian grid, and less than 0.001% for the standard cubic grid (Figure 4). The latter is found to be equally beneficial in longer climate simulations (1 year) without activating the mass fixer (not shown).

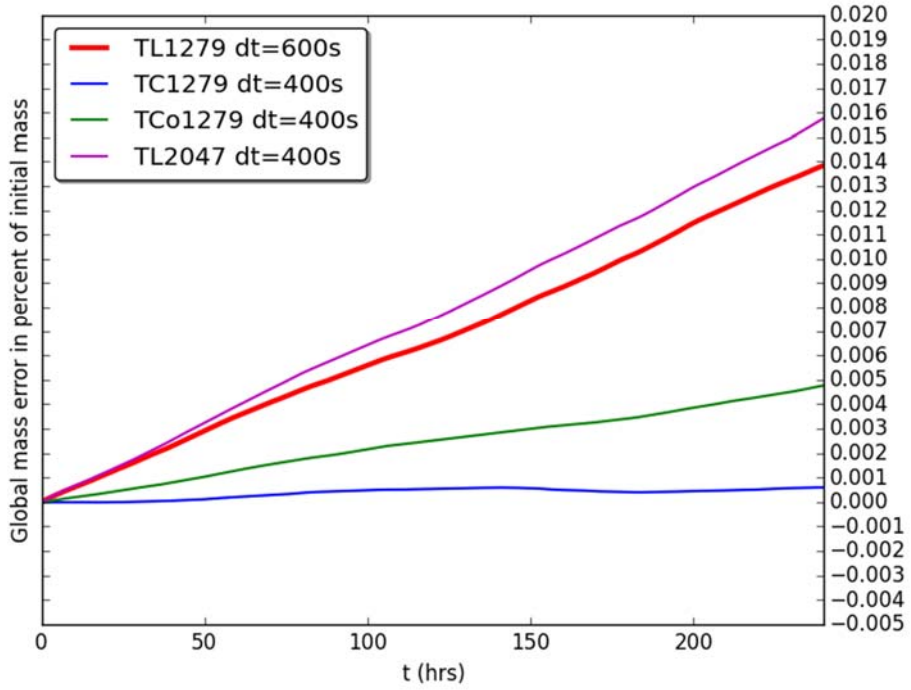

Figure 4: Global mass conservation error for three forecasts at 1279 spectral resolution: linear reduced Gaussian grid with 600s timestep (red), (ii) TC1279 cubic reduced Gaussian grid with 400s timestep (blue) and the (iii) TCo1279 octahedral reduced Gaussian grid with 400s timestep (green). For comparison, also the TL2047 is shown (violet). The chosen time steps are optimized for meteorological accuracy and efficiency at the given resolution.

3 Review of developments at other NWP centres

Several NWP centres are following a *disruptive approach* where the model is entirely reformulated. Mostly, the vertical coordinate in use is no longer pressure but height based. The horizontal discretization is quasi-uniform in most areas but does not preserve the global latitude-longitude structure anymore, and does not allow for spectral transforms. Representative examples of models with characteristically different horizontal discretization are summarized in Table 1. A more comprehensive list of existing models for the atmosphere can also be found in Marras et al. (2015).

Several national weather services have already initiated long-term programmes to change their NWP models in coordination with future HPC infrastructure environments:

- Following the High Performance and High Productivity Computing (HP2C) project, MeteoSwiss is already running daily real-time suites of COSMO-1 (the future ultra-high resolution 1.1 km model of MeteoSwiss) on a Cray XK7 equipped with NVIDIA Tesla K20x processors with a version that has been fully adapted to hybrid architectures (Fuhrer et al., 2014).
- The Met Office has initiated a long-term programme called LFRic (after L.F. Richardson) to review their entire software and hardware infrastructure, and to advance the existing model dynamical core (ENDGAME) following developments under the nationally funded NERC project Gung-Ho.
- The German weather service (DWD) in collaboration with the Max-Planck Institute for Meteorology (MPI-M) has developed a new model (ICON) in view of the scalability challenges and recently replaced GME with the ICON model for their global forecasts.
- The RIKEN/Advanced Institute of Computational Science in Japan has initiated a post K-computer project to develop exascale capabilities for climate simulations based on accelerators.
- Environment Canada has upgraded their NWP model recently based on a Yin-Yang grid configuration using two overlapping latitude-longitude grids. Nevertheless, more scalable solutions are being envisaged.
- The Japanese Meteorological Agency (JMA) is in a similar position as ECMWF with an efficient global hydrostatic spectral transform model. They are investigating different options for the future including the NICAM model, double Fourier series for the transforms in the existing global spectral model, and/or the use of their limited area finite-volume model (ASUCA) for the global domain (Ishida et al., 2010).

- The G8 funded project ICOMEX aimed at comparing several icosahedral grid models and address common scalability challenges (ICON, NICAM, MPAS; see also Table 1).
- Several groups in the US are testing candidate future models as part of the Next Generation Global Prediction System (NGGPS) programme for NOAA's National Weather Service (NWS), with a view to increase the competitiveness of NOAA's weather prediction capabilities and scalability on future HPC platforms (GFS, NIM, MPAS, NUMA/NEPTUNE, FV-3, NMMB-UJ; see also Table 1). There is a recommendation to take forward MPAS and FV-3 which are both based on finite-volume discretization.

Type	*Model/Project name*	*Reference*
Yin-Yang grid	GEM	Qaddouri and Lee, 2011
Continuous (CG) or discontinuous Galerkin (DG)	NUMA/NEPTUNE	Giraldo et al., 2013, Marras et al, 2015
Spectral element	CAM-SE	Dennis et al., 2012
mixed finite-elements	Gung-Ho	Cotter and Shipton, 2012; Thuburn and Cotter, 2015
finite-volume, cubed sphere	GFDL FV3; NMMB-UJ	Putman, 2007; Janjic and Gall, 2012
finite-volume, hexagonal icosahedral	NICAM	Satoh et al., 2014
finite-volume, hexagonal icosahedral	MPAS; ICON-IAP; NIM	Skamarock et al., 2012; Gassmann, 2013; Lee and McDonald, 2009
finite-volume, triangular icosahedral	ICON-DWD	Zängl et al., 2015
finite-volume, fully unstructured	PantaRhei/FVM (cf. section 5)	Szmelter and Smolarkiewicz, 2010; Smolarkiewicz et al., 2013

Table 1: Summary of models with characteristically different horizontal discretisation.

So far, no preferred choice has emerged as an obvious alternative since all options come with advantages and drawbacks, mostly efficiency issues on existing architectures, but also numerical accuracy issues such as grid imprinting at “special points” and/or the need for excessive orography smoothing and dissipative filters. Illustrating this point, the Korean Meteorological Agency (KMA) started a 9-year programme in 2011 to achieve world-class global NWP status by 2019 (KIAPS). The development of a new dynamical core is a central component in this undertaking and interestingly, KMA still appear to pursue almost all of the above (semi-structured) options including double Fourier series spectral transforms (Park et al., 2013) four years into the project (Park, 2014).

Figure 5 illustrates the relative scalability of several candidates in Table 1 from the scalability intercomparison forming part of the NGPPS programme (Michalakes et al., 2015). The figure expresses the number of cores required for each of the models to reach the operational target of simulating 7

forecast days in one hour (the red dashed line). This is weaker than ECMWF's present requirement of 10 days in one hour. With increasing numbers of cores, the figure also illustrates how the different models exceed that requirement expressed in multiples of the threshold. From these results the hydrostatic IFS (green curve) emerges as the most efficient operational global NWP model compared to the new generation non-hydrostatic models, being the fastest model in time-to-solution and requiring the least number of cores to do so; both at the tested 13km and the 3km target resolutions. Notwithstanding, that the hydrostatic balance assumption underlying the hydrostatic primitive equations will become invalid as horizontal grid spacings reach O(1km), global non-hydrostatic model formulations combined with novel local discretization are evidently still more costly. Moreover, none of the models reach the 7 forecast days per day target at 3km resolution (Figure 6). However, in terms of scaling efficiency IFS is the worst (not shown), due to the communication overhead already mentioned earlier. Hence there is no question that this status can only be maintained with a successful transition to an energy-efficient global model operating at non-hydrostatic scales, and by operating on the emerging (and as yet unavailable) computing architectures of the future.

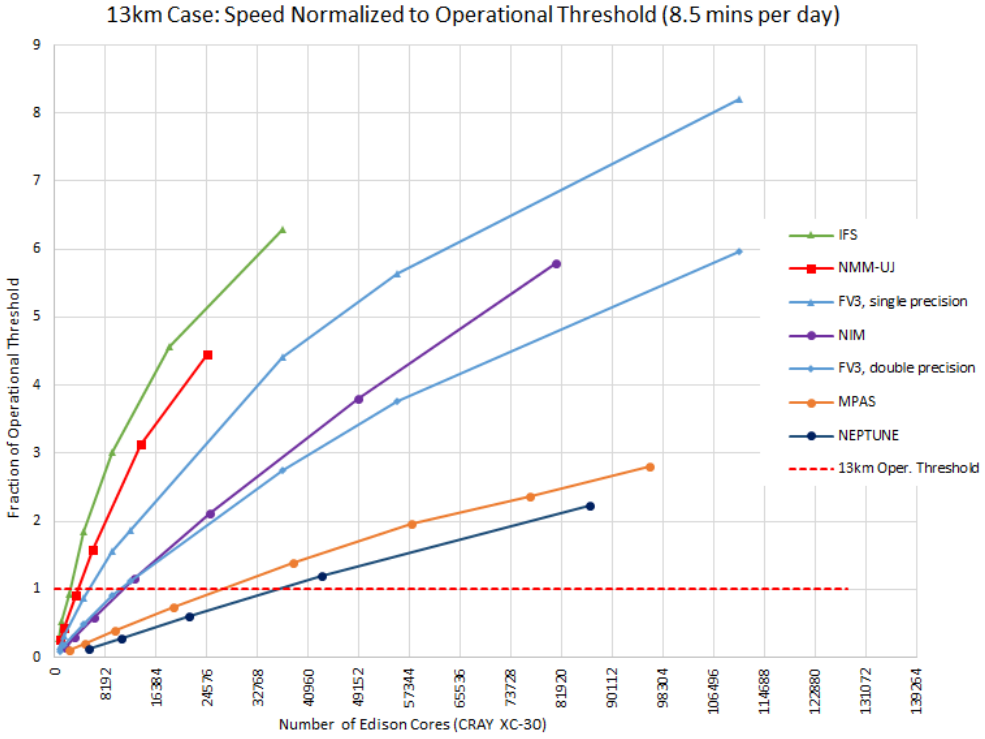

Figure 5: NGGPS 13 km case showing the *time-to-solution* as a function of Cray XC-30 cores for a number of candidate U.S. models (all non-hydrostatic) and in addition IFS (hydrostatic, green) as a guest model. All models were configured to run a baroclinic wave test case (no physics), with the addition of 10 3D tracers. Performance is shown relative to an NGGPS operational threshold of 8.5 minutes per day, where the candidate dynamical cores are assumed to take 50 percent of the time.

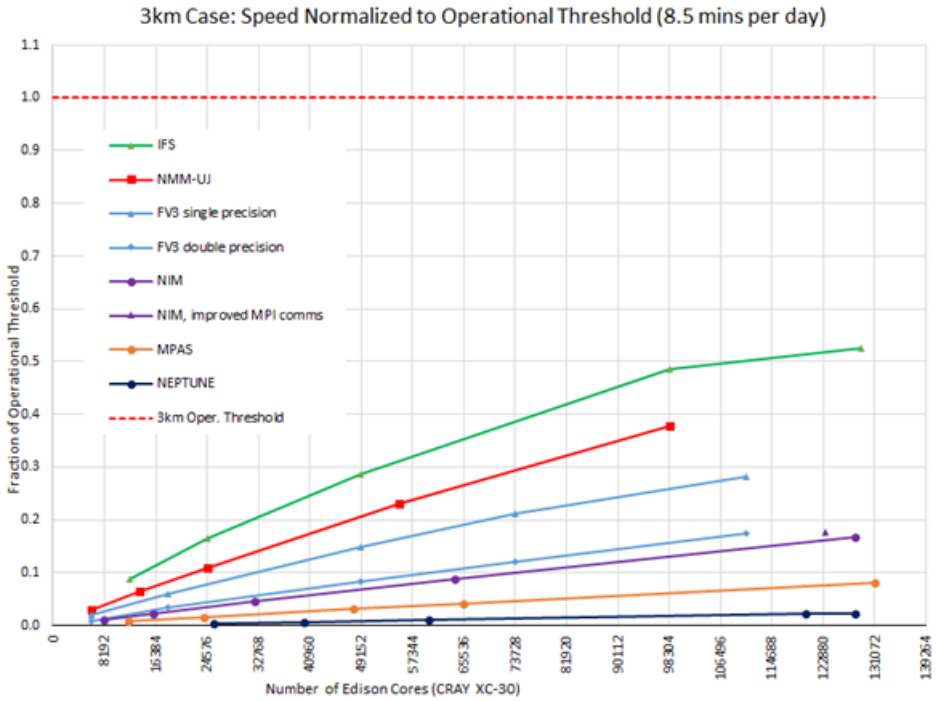

Figure 6: as Figure 5 but for the NGGPS 3 km case.

4 Roadmap for development

ECMWF's objective is to develop one of the most advanced and flexible model infrastructures for operational, global NWP applications to optimally exploit future computer hardware emerging over the next 20 years. In addition, the code infrastructure should facilitate efficient collaboration with member states, both for scientific exploitation and the effective use of boundary conditions from ECMWF for limited-area modelling.

A number of key areas are addressed in the 2015-2018 timeframe where the current IFS does not allow sufficient flexibility and hence hinders progress in adapting to future computing architectures. In order to increase substantially the readiness level for emerging technologies, we seek alternative numerical algorithms that depend on different communication patterns, both local and global. It is likely that several different numerical algorithms will be justifiable and efficient in terms of *time-to-solution* but that they will not be competitive in terms of *energy-to-solution*. The latter is emerging as a growing problem, with rising energy costs absorbing substantial parts of the compute budget in weather prediction services. Today's NWP models achieve less than 10% of peak sustained performance or less. Yet despite the low compute efficiency, today's energy consumption does not scale with the compute, and a substantially higher percentage of the energy usage peak is consumed. Thus apart from increasing computational density in NWP models, which increases compute efficiency, the use of emerging technology that consumes energy according to the compute rate must be considered.

The scope of the activities on numerical methods and technical model development over the next three years (some activities are externally funded) is to fundamentally establish flexibility, new programmatic interfaces and data structures, together with exploring scientific choices which open up new possibilities

for a complete description of the Earth-System, which scale better and use emerging future hardware in an energy-efficient manner.

By developing a hybrid and hierarchical methodology that continues to support the spectral transform technique while at the same time (and optionally but not necessarily on the same grid) supporting a disruptive technology, the development risk is substantially minimized, while introducing the required flexibility when needed. Considering that the resolutions used in Figure 5 and Figure 6 approximate potential operational target resolutions for 2020-2030, respectively, it clearly supports this strategy, given the current competitive advantage of IFS in terms of time-to-solution on existing state-of-the-art hardware. Moreover, the technical possibilities resulting from research into very high-speed networks and optical processing units, and the potentially saturating number of MPI tasks (cf. table 2) could make spectral transforms competitive well beyond the 10 year horizon.

Moreover, the numerical techniques explored facilitate adaptation to the nature of the simulated information, namely multi-scale (both in time and space) phenomena, from-global-to-local-to-global scale interactions, and increasing complexity of the simulated processes. The stiffness of the problem increases substantially with the presence of global and local scales and increasing degrees of freedom, which translates into an increased computational cost. However, better exploiting the knowledge of global and local interactions may provide substantial efficiency gains if for example limits of accuracy or sufficiency of representation can be clearly established for each simulated process.

One strategic goal by 2025 is to establish the efficiency required to run at least twice daily 51 member ensembles at 5km globally uniform resolution. This goal is achievable considering projected technological advances of computing architectures, but requires substantial advances in the numerical methods used and their adaptation to new technologies in order to achieve a reasonable energy-to-solution. Scientific advances required include an effective physics-dynamics coupling in this regime (section 5.7).

In recognition of the development need, ECMWF has not only invested substantial effort in optimizing the current model for near-future resolution and complexity upgrades, but also explores novel scientific and numerical options. In support of this research, ECMWF is leading a project funded by the European Commission's Horizon 2020 funding framework, Energy-efficient Scalable Algorithms for Weather prediction at Exascale (ESCAPE), in collaboration with 11 partners from member states, industry and academia, to develop Energy-efficient, SCalable Algorithms for weather Prediction at Exascale, Moreover, ECMWF is hosting the European Research Council (ERC) funded, frontier research project PantaRhei (FP7/2012/320375). PantaRhei aims to prepare the mathematical and numerical technology for the next generation of weather and climate models, which, paired with the scalability programme efforts, will form the foundation of the future IFS.

The technical foundations necessary to introduce the required flexibility on the algorithmic choices and on the mapping to different computing architectures is described in section 5.1. Following a hybrid methodology, sections 5.2 and 5.4 describe the implementation of a particular finite-volume non-hydrostatic module within the IFS framework, while admitting other more disruptive horizontal discretization possibilities. Other sections address particular focus areas identified for further development, which are equations (section 5.3), the time discretization and flow solvers (section 5.4), the transport of many species (section 5.5), the vertical discretization including the treatment of

orography (section 5.6), the coupling to physical parameterizations and diabatic processes with particular emphasis on partially resolved processes (section 5.7), support for the development of tangent linear and adjoint models (section 5.8), quantifying model and forecast uncertainty (section 5.9), and aspects to be considered for the coupling and interaction with other components of the Earth-System such as the ocean (section 5.10). All these aspects of the roadmap are illustrated in Figure 7 together with a time-line of developments, with major milestones anticipated for 2020 and 2025, respectively. The yellow background colour indicates more certain developments and the green background colour indicates developments that are less certain due to the uncertainty of the underlying and evolving hardware technology or because there are unsolved scientific problems. Table 2 lists potential targets for operational implementation and illustrates a trend towards more local shared memory parallelism embedded into a constant or more slowly growing task parallelism. The different options indicate considerable uncertainty with respect to the distribution of tasks and threads. Only the ensemble configuration is listed since this is the dominant cost driver for current and future operational implementations.

Year	Resolution	Parallel configuration per member	Cost factor
2016	51 x TCo639 (~18km)	160 tasks x 12 threads	1
2020	51 x TCo1279 (~9km)	1600 tasks x 12 threads 160 tasks x 120 threads	10
2025	51 x TCo1999 (~5km)	1600 tasks x 120 threads 160 tasks x 1200 threads	70

Table 2: Target implementations and corresponding computational cost for the ensemble system, which is the most cost-significant part of the operational forecast suite. The parallel configuration indicates the observed trend to more local shared memory parallelism that needs to be exploited.

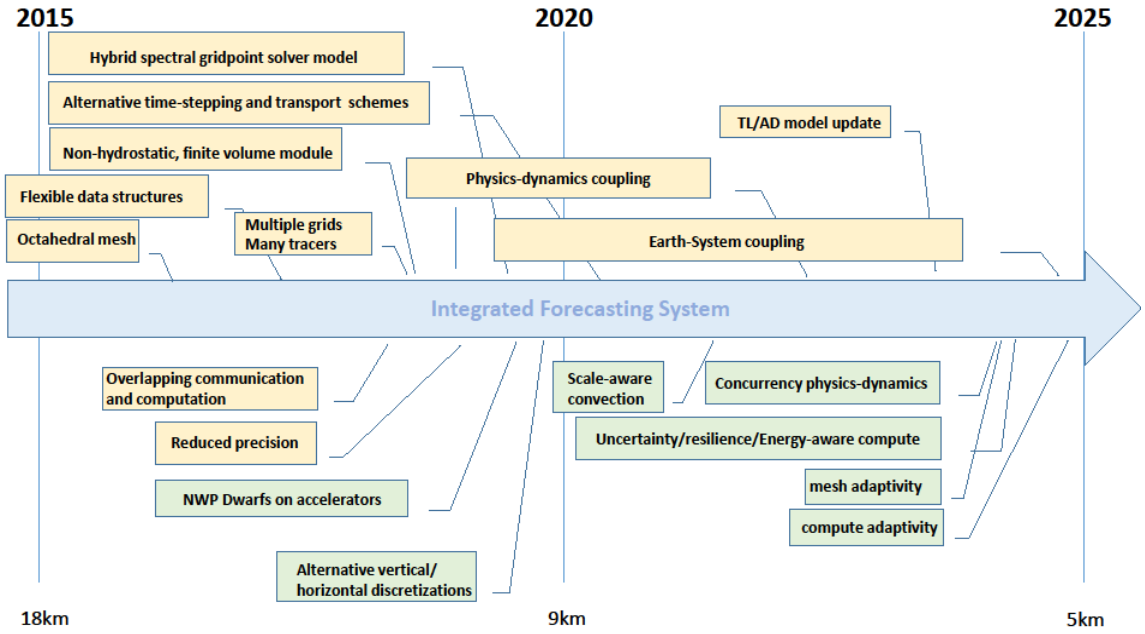

Figure 7: Roadmap of development and potential targets for the period 2015 -2025. Most targets are explained in section 5, reduced precision refers to selected computations or communications using less bits, NWP Dwarfs encapsulate a (performance) relevant characteristic or required functionality of a NWP or climate model that is packaged as a runnable and verifiable mini-application.

5 A flexible, scalable, and sustainable model infrastructure

5.1 Scalability, novel data structures and technical design

In order for ECMWF's applications to optimally exploit future computer hardware emerging over the next 20-30 years, a flexible and dynamic data structure framework, named *Atlas* is being developed to serve as a foundation for a wide variety of applications, ranging from the use within the IFS model for the development of alternative dynamical core modules (cf sections 5.3 and 5.4), to current and future developments in the forecast department and under the auspices of the scalability programme (cf. *Hermes project*). A particular aspect is the much needed replacement of the single-threaded *emoslib*, a library of tools and algorithms that provides ECMWF's Meteorological Archival and Retrieval System (*MARS)* and the visualisation and data processing software *Metview* with all the fundamental data processing capabilities, all of which are used for the daily product generation for the member states.

The *Atlas* framework provides parallel distributed, flexible, object-oriented data structures for both structured grids and unstructured meshes on the sphere. It separates concerns of mathematical model formulation and numerical solutions from the cumbersome management of unstructured meshes, distributed memory parallelism, and input/output of data. It is recognised that handling flexible object-oriented structures and carefully controlled memory-management, as would be required with the expected deepening of memory hierarchies in future hardware, is not easily achieved with the Fortran language. Hence, the language of choice for *Atlas* is C++, a high performance language providing excellent object-orientation support, and building upon C's memory management proficiency. A

Fortran2003 interface exports all of *Atlas*' functionality to Fortran applications, hence seamlessly introducing modular object-oriented concepts to IFS. Moreover, many other C++ based ECMWF applications can directly benefit from *Atlas*.

Technically, the application (e.g. IFS) instructs *Atlas to* generate a Mesh. This object stores the horizontal coordinates of every node and requires connectivity tables between the nodes via elements such as triangles, quadrilaterals and lines, as is necessary for unstructured grids. As the coordinates and connectivity tables can have a large memory footprint for large meshes, the *Mesh* is a distributed object, meaning that the mesh is subdivided in partitions and each parallel task is responsible for one partition. Using the *Mesh, FunctionSpace* objects can be created on demand. A *FunctionSpace* describes in which manner *Fields* are discretized on the mesh. A straightforward *FunctionSpace* is the one where fields are discretised in the nodes of the elements. Other *FunctionSpace* objects could describe fields discretised in cell-centres of triangles and quadrilaterals, or in edge-centres of these elements. Another type of *FunctionSpace* describes spectral fields in terms of spherical harmonics. *Fields* are objects *that* store the actual data contiguously in memory as a one-dimensional array and can be mapped to an arbitrary indexing mechanism to cater for e.g. the current IFS memory layout, or a different memory layout that proves to be beneficial on emerging computer hardware. *Fields* are addressed by user-defined names and can be associated to *Metadata* objects, which store associative information like the units of the field, grib parameter ID, or a time-stamp. It is this flexibility and object-oriented design that leads to more maintainable and future-proof code. Figure 8 summarises the aforementioned objects and their connections in the object-oriented design together with the relevant classes.

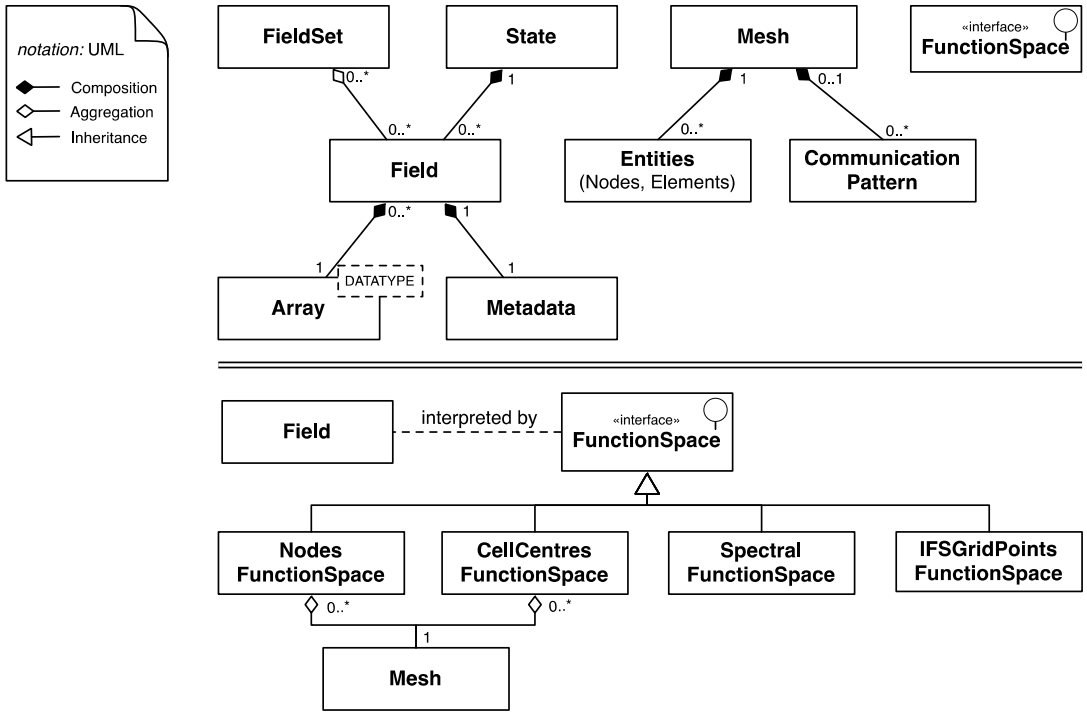

Figure 8: Atlas data structure components and their connections as described in the text.

The MPI parallelisation relies on the distribution of the computational mesh with its unstructured horizontal index. The *Atlas* framework is responsible for generating a distributed mesh, and provides communication patterns using MPI to exchange information between the different partitions of the mesh.

To minimise the cost of sending and receiving data, the distribution of the mesh is based on an equal-regions domain decomposition algorithm optimal for a quasi-uniform node distribution on the sphere (Leopardi, 2006; Mozdzynski, 2007). The equal-regions domain decomposition divides the sphere into bands oriented in the zonal direction, and subdivides each band in a number of regions so that globally each region has the same number of nodes. Notably, the bands covering the poles are not subdivided, (communication-wise) forming two polar caps. Figure 9 illustrates the domain decomposition with 1600 equal regions distributed over a corresponding number of MPI tasks, as applied in the TCo1279 octahedral reduced Gaussian grid planned for the next horizontal resolution upgrade (see also section 5.2).

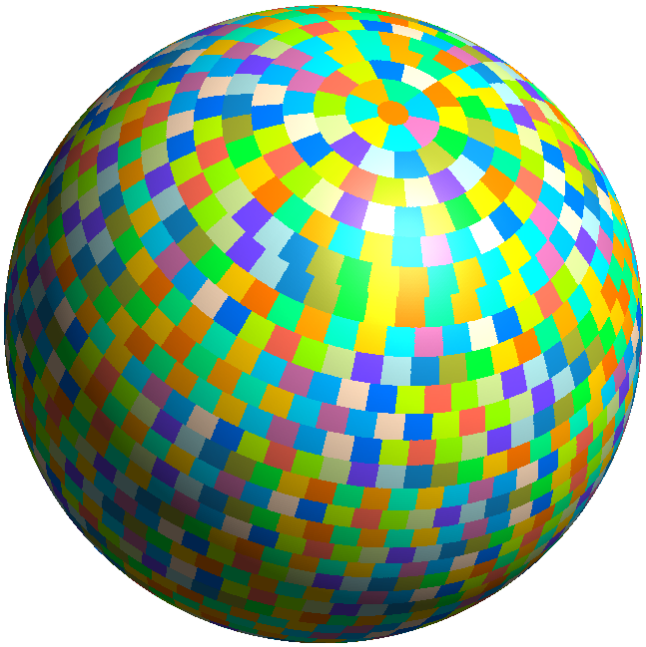

Figure 9: Equal-regions domain decomposition with 1600 partitions as typically applied in the new TCo1279 octahedral reduced Gaussian grid planned for the next horizontal resolution upgrade.

Because global communication across the entire supercomputer cluster will become prohibitively expensive, numerical algorithms may be required to limit communication to --- even physically located --- nearest neighbours. Such communication typically happens through thin halos surrounding every mesh partition, creating an overlap with a neighbouring partition. *Atlas* provides routines that expand the mesh partitions with these halos and provides communication patterns to update fields in the halos with values from neighbouring partitions.

With communication patterns of applications changing from few large messages to many short messages, computer network latency becomes relatively more important and intelligent data placement to shorten message pathways and to reduce network congestion, together with a consideration of the overall HPC network topology matter. So called rank-mapping algorithms create a one-to-one mapping between the MPI ranks of the application and the physical compute nodes, for example developed for the NICAM model on the K computer (Kodama et al., 2014). There exist two active areas of research, to find appropriate HPC topologies that may obviate the need for explicit mappings between an

application and compute nodes, or to design dynamically configurable interconnections considering the specific application needs.

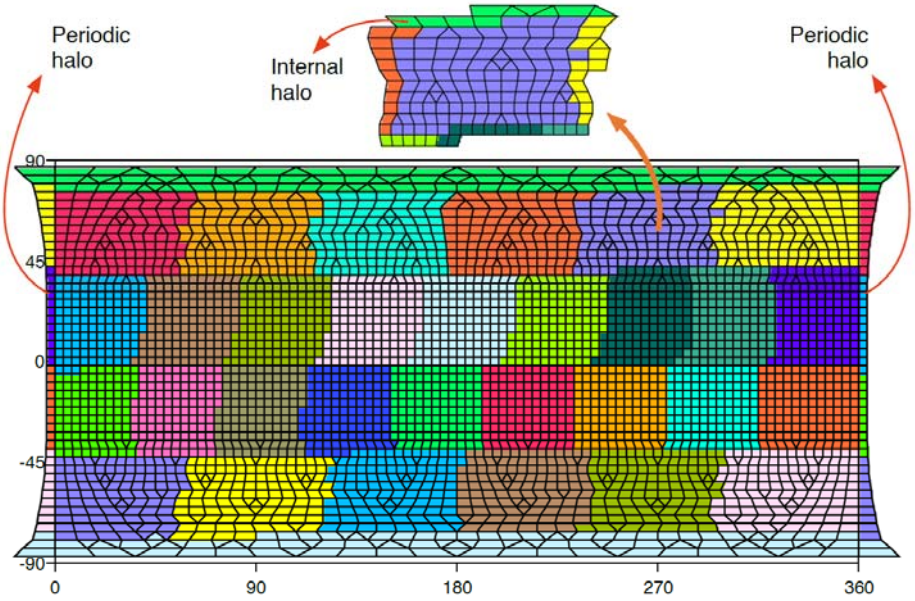

Figure 10: Equal-regions domain decomposition for a N24 reduced Gaussian grid. Also shown is the internal halo of one partition and the periodic halos for periodic boundary conditions (i.e. reconnection on the sphere).

Figure 10 shows the domain decomposition for a standard N24 reduced Gaussian grid in 32 partitions, and illustrates the creation of internal and periodic halos. With periodic halos the periodic boundary is treated exactly like any other internal boundary between different partitions. The partitions involving the poles are periodic with themselves.

5.2 Horizontal discretization

The IFS relies on spectral transforms and corresponding computations in spectral space to compute horizontal derivatives valid at the nodal points of the Gaussian grid. The computation of these derivatives comes at a substantial cost due to the communication overhead (Figure 2) and hence the use of derivatives has been minimized carefully (Wedi et al., 2013).

Pending advances in communication hiding technologies, alternative horizontal discretisation algorithms are required that minimize communication cost while relying on compact stencil gridpoint computations that only require local (nearest-neighbour) communication. Such discretisation methods are typically based on meshes composed of elements such as triangles, quadrilaterals, or lines. Examples include the finite volume method (FV), the finite element method (FE), or higher-order (>2) methods such as the Discontinuous Galerkin method (DG) and the Spectral Difference method (SD).

Although these discretisation methods can be used on arbitrary unstructured meshes, we explore a hybrid approach, where compatibility with a principle latitude/longitude nodal structure as provided by the reduced Gaussian grid is maintained and thus spectral transforms may still be used. Bespoke unstructured meshes can be generated that include the reduced Gaussian gridpoints. By sharing the same data points and the same domain decomposition as is currently used by IFS, an evolutionary and

seamless transition is guaranteed while offering new opportunities for principally local computations. In particular, a cheaper option to compute local derivatives opens new avenues of research previously unavailable due to technical and efficiency constraints.

As part of the PantaRhei project, a FV dynamical core has been developed and tested with bespoke meshes based on the standard reduced Gaussian grids. As one outcome of these investigations, it was evident that discretisation errors can be diminished significantly by modifying the reduced Gaussian grid. The new design is derived from triangulating a regular octahedron and projecting it on the sphere, while keeping latitudes fixed at the standard Gaussian quadrature nodes defined by the roots of the ordinary Legendre polynomials. Figure 11 illustrates the meshing of a coarse N24 reduced Gaussian grid with an approximate resolution of 3.75 degrees. The more uniform triangulation of the octahedral mesh (Figure 11 right panel) has a more uniform local resolution, as illustrated by the shading, than the standard mesh (Figure 11 left panel). The more locally uniform stencils of the octahedral mesh lead to a better cancellation of numerical discretisation errors in FV computations and the number of points for the same spectral truncation is further reduced (cf section 2.1).

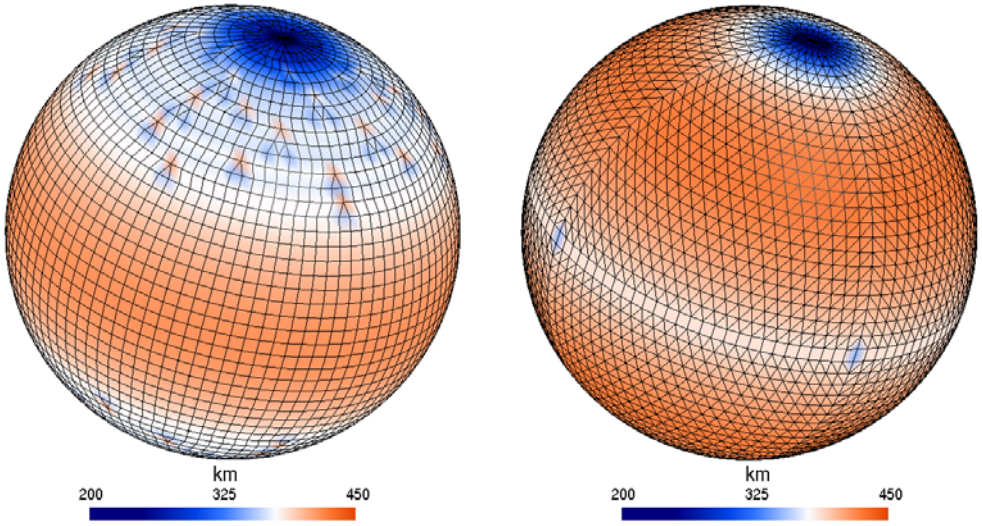

Figure 11: Primary meshes generated around the N24 reduced Gaussian gridpoints with 3.75 degrees (or approximate grid distances of 416 km). The shading represents the effective grid distance of the median-dual mesh, which is the mesh used for the FV integrations, derived as the square root of the local median-dual mesh area size. The octahedral mesh shows a locally more uniform dual mesh resolution than the standard mesh.

In the future, the gridpoint horizontal discretisation offers the possibility to efficiently create a set of hierarchical meshes which will be used for multigrid preconditioning of the elliptic problems arising from the semi-implicit time discretization of the system of equations, or to facilitate alternative time-stepping algorithms (section 5.4). Indeed, alternative quadratures on the sphere exist for the octahedral setup that would alter the position of the latitudes but could provide more accuracy and efficiency in the context of hierarchical multigrid arrangements while maintaining the option to use spectral transforms (van Lent, 2015). Moreover, explicitly created coarser meshes may be used to represent selected physical processes, e.g. the radiation model is computed today on a different reduced grid but using exactly the same domain decomposition. Finally, hierarchical meshes minimize communication and

redistribution of points but could be used to generate added information in regions of interest, e.g. wave and ocean models may benefit from a refined representation of coastlines, or air quality chemical modelling can make use of fine-scale emission data.

5.3 Equations

The simulated vertical extent of the atmosphere is relatively thin compared to its horizontal extent, and in particular vertically propagating sound waves, admitted by the fully compressible Euler equations, impose severe restrictions on the numerical algorithms used. Hydrostatic primitive equations (HPEs) filter vertically propagating sound waves by virtue of the hydrostatic approximation, facilitating larger time-steps in the numerical integration. HPEs currently form the basis of the IFS in all operational forecasting applications at ECMWF. A central aspect of the HPEs is the straightforward separability of the horizontal and vertical discretisation, which facilitates the design of highly efficient and robust SISL integration schemes as used in the IFS.

Notwithstanding, the hydrostatic balance assumption underlying the HPEs will become invalid as horizontal grid spacings reach O(1 km). This may occur gradually by means of a uniform global resolution increase, or much earlier through local grid refinement in high-resolution, convection-permitting modules coupled to the HPE model. In either case, future global NWP applications require non-hydrostatic formulations, and this task continues to stimulate the discussion about suitable alternative and potentially more efficient subsets of the compressible Euler equations (e.g. Ogura and Phillips, 1962; Lipps and Hemler, 1982; Durran, 1989, 2008; Davis et al., 2004; Arakawa and Konor, 2008; Klein, 2009; Dubos and Voitus, 2014). A common motivation behind alternative non-hydrostatic equations is to neglect terms responsible for meteorologically insignificant (acoustic) modes, while retaining all terms required to represent the relevant modes for NWP without distortion.

The Arakawa-Konor (AK) “unified” system of the governing equations (Arakawa and Konor, 2008) combines properties of sound-proof non-hydrostatic equations at small scales and the compressible HPEs at large hydrostatic scales. Therefore, the AK system represents a viable all-scale set of governing equations for global NWP. Especially, when cast in the eta-p vertical coordinate framework of the IFS, the AK system shows similarities to the HPE formulation. This triggered an investigation into the AK system as an alternative non-hydrostatic set of governing equations for the IFS that could potentially resolve some of the problems seen with the non-hydrostatic model (NH-IFS) implementation based on the compressible Euler equations. In a collaborative effort with Météo-France, a numerical solution approach for the AK system with the spectral SISL discretisation as used in the IFS has been developed (Voitus et al., 2015). However, the main conclusion from this research is that, even though the AK equations have useful analytical properties, they appear less efficient to solve than the compressible Euler equations in the framework considered. In particular, while the solution of compressible Euler equations with the SI scheme relies solely on the efficient direct spectral inversion of the underlying Helmholtz problem, the AK equations demand an additional inversion of a three-dimensional Poisson problem that adds significantly to the computational cost and to the general complexity of the model formulation. The experience gained so far has led to the decision to abandon the efforts regarding the AK equations and to focus on improving the existing NH-IFS as part of the actions in the H2020 project ESCAPE, and to pursue alternative solution procedures for the compressible Euler equations.

The high efficiency, robustness and proven fidelity of the hydrostatic model (H-IFS) makes it beneficial to maintain this model configuration as long as possible within its regime of validity. Considering the strategy for an ensemble at 5km uniform resolution in 2025, the use of H-IFS may still be viable and competitive. However, a consistent all-scale non-hydrostatic dynamical core used throughout the operational systems is desirable for the longer term future. In the short to medium-term, this may be realised by an improved NH-IFS based on the spectral SISL discretisation. In the mid- to long-term, the anticipated evolution of computational architectures and the changing requirements on numerical methods with explicitly simulated convection and beyond may make H-IFS redundant, and consequently lead to a single seamless all-scale dynamical core for future global NWP.

In order to account for the gradual demise of the hydrostatic model without loss of efficiency, PantaRhei has followed a hybrid approach by developing a complementary non-hydrostatic all-scale finite-volume module (FVM). The development started from an implementation based on the soundproof anelastic system that was progressively extended to the pseudo-incompressible and compressible Euler equations. Distinct elements of soundproof models as building blocks of more general compressible frameworks include the accurate treatment of boundaries and semi-implicit time-integration using an effective 3D elliptic solver.

The result is a consistent discrete global finite-volume modelling framework for large-time-step integration of the non-hydrostatic deep-atmosphere anelastic, pseudo-incompressible, and compressible Euler equations (Smolarkiewicz et al., 2014; Smolarkiewicz et al., 2015). Using FVM has enabled a rigorous study of the various governing equations and their particular solution approaches with regard to future requirements. Based on this study, our current understanding is that in terms of the regime of validity (Smolarkiewicz et al., 2014; Kurowski et al., 2015) and in terms of solution efficiency, the compressible Euler equations are favoured for global non-hydrostatic NWP. However, other applications, such as data assimilation, ocean modelling, or simulations of solar convection may benefit from asymptotic limit solutions provided by reduced soundproof equations.

5.4 Time-discretization and flow solvers

When characterising computational fluid models, it is common to specify both the spatial and the temporal discretisation of the governing partial differential equations (PDEs). Table 1s 1 and 2 in Marras et al. (2015) are representative examples for NWP forecasting systems and atmospheric research models, respectively. The abundance of spatial discretization techniques (see Table 1 in section 3 for representative examples) combines with a variety of “time-stepping scheme” choices, creating a multitude of numerical methods for integrating atmospheric PDEs.

Efficient time-stepping is one of the most important aspects for time-critical NWP and substantially contributes to the current advantage for SISL methods. Moreover, implicit and semi-implicit schemes for the numerical integration of non-linear dynamical systems have several decisive conceptual advantages over explicit schemes, including superior conservation properties, robustness for arbitrary time-step sizes and improved adherence to dominant physical balances. But implicit schemes require more communication in MPI implementations, so that their advantages may be offset by diminishing efficiency on massively parallel future architectures. Although Figure 5 and Figure 6 suggest that this is not the case with IFS for existing or near-future hardware, this is a focus area for mathematical research and including parallelism in time to overcome the restrictions imposed by explicit, sequential time-

stepping (e.g. Horten and Vandewalle, 1995). The technical framework described in sections 5.1 and 5.2, together with the adaptive numerical algorithms currently developed, will facilitate the practical exploitation of space-time congruence.

Given its importance, the following covers in some more detail the development of the all-scale finite-volume module (FVM; Smolarkiewicz et al., 2015), newly developed under the auspices of the ERC Project PantaRhei hosted by ECMWF, including the development of a powerful bespoke 3D elliptic solver and involving entirely different communication and computational patterns compared to SISL.

There are two distinct, high-level categories of time-stepping schemes. One is the method of lines (MOL), where spatial and temporal discretisation are viewed as independent from each other. This allows all spatially discretized terms of the governing PDEs to be placed on the right-hand-side (rhs) and solving the resulting prognostic system of equations, symbolized here as $d\psi/dt=R(\psi(t), t)$, in the spirit of ordinary differential equations (ODEs) to a desired accuracy in time, at all spatial locations of the discrete model domain. Consider that R may depend on ψ values in many neighbouring, or even all discrete spatial locations. This technique is popular in many computational fluid dynamics applications, including atmospheric, oceanic and air-quality models. Representative examples include basic centred in time and space methods or multi-stage Runge-Kutta temporal integrals of centrally (e.g. pseudo-spectrally) discretized PDEs (Jameson, Schmidt and Turkel, 1981). Such schemes may come in different flavours of bespoke mixed explicit-implicit time integrations of various terms on the rhs, including iterative sub-stepping of selected terms in a subclass of split-explicit schemes (Skamarock and Klemp, 2008; Satoh et al., 2008). Computational stability of MOL schemes depends on the magnitude of the inverse time scales $|dR/d\psi|$ of the processes combined in R, including the spatially discretised nonlinear advection operator (convective derivative) $\mathbf{v}.\nabla\psi$. The latter limits the integration time step δt by the time scale of advection $\delta x /|\mathbf{v}|$ where δx marks the smallest distance in any direction between the discrete locations in the model domain.

An alternative approach to the method of lines is a class of algorithms integrating the governing PDEs over an element of the 4D time-space continuum, effectively viewing the governing equations as physical constraints of arbitrary path or volume integrals of differential forms in the 4D space (cf. Smolarkiewicz and Pudykiewicz, 1992). A prominent example are the SL schemes that revolutionised NWP two decades ago. SL schemes generate solutions to governing PDEs $\partial\psi/\partial t+\mathbf{v}.\nabla\psi=R \Leftrightarrow D\psi/Dt=R$ as trajectory integrals $\psi(x, t) = \psi(\mathbf{x_o}, t_o) + \int_T R d\tau$; see sketch in Figure 12 for an illustration. Consistent with principles of Newtonian mechanics, and in contrast with MOL, the convective derivative is absorbed in the path derivative $D\psi/Dt$ on the lhs.

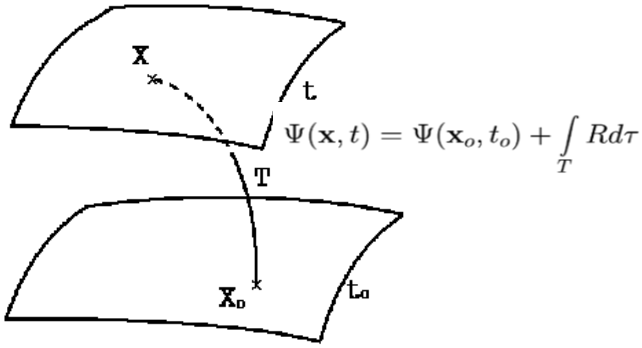

Figure 12: Sketch illustrating the trajectory integral.

For forcing R≡0, any specific fluid property ψ is translated along flow trajectory T without change. In terms of numerical approximations, this results in unconditional stability of SL advection schemes, a desirable property for time-critical NWP. For R≢0, the trajectory integral of R can be approximated using a weighted average of the R values from physical space at a past time t_0, where the state of the atmosphere is known, and the values from the space at a future time t where the solution is sought. When the approximation solely relies on the past values, the integration scheme is explicit. In particular, when R is constant in time and space, $\psi(x, t) = \psi(\mathbf{x_o}, t_o) + R\,\delta t$; where $\delta t = t - t_o$. Generally, however, R depends on ψ, its derivatives, and the position in time and space. The time interval δt then has to be sufficiently small to resolve the variation of R along T between x_o and x, as otherwise the amplitude of the approximate solution may grow unboundedly. However, if the approximation depends on the future (unknown) values of R the boundedness of the solution can be assured, but the resulting scheme becomes implicit. For the system of atmospheric PDEs, implicit integration schemes lead to intricate diagnostic problems in form of elliptic boundary value problems (BVPs), the solutions of which typically provide future pressure perturbations sufficient to balance the explicit part of these solutions --- $\psi(x_o, t_o)$ and $R(\psi(\mathbf{x}_o, t_o), \mathbf{x_o}, t_o)$ in our sketch --- in accord with the properties of the governing PDEs and the imposed boundary conditions. Although these BVPs can be intricate and difficult to solve (cf. Smolarkiewicz et al., 2014), implicit integration schemes lead to superior solutions not only in terms of computational stability (and thus efficiency due to the longer time-steps) but also in terms of the accuracy (Knoll et al., 2003; Wedi and Smolarkiewicz, 2006; Dörnbrack et al., 2005). In particular, SISL schemes have been at the heart of successful NWP, and of the IFS in particular (cf. section 2.1). Although in principle MOL schemes that are implicit with respect to the advection could be considered, in practice they are seldom used. The nonlinearity of advection and its hyperbolic (initial value) nature adversely affect solvability of the resulting BVPs, and such implicit schemes are inaccurate for δt larger than admitted for explicit schemes (section 1.3 in Hundsdorfer and Verver, 2003). In contrast, accuracy of SL schemes does not depend on the absolute magnitude of the flow velocity, but on the magnitude of its derivatives (Smolarkiewicz and Pudykiewicz, 1992), thus providing excellent accuracy for relatively smooth flows.

The scheme favoured in FVM which is akin to SL and distinguished from MOL are forward-in-time finite-volume (FTFV) integrators of the flux-form archetype model $\partial\psi/\partial t + \mathbf{v}.\nabla\psi = R \Leftrightarrow D\psi/Dt = R \Leftrightarrow \partial\rho\psi/\partial t + \nabla.\rho\mathbf{v}\psi = \rho R$, where ρ is the fluid density. Notably, for any constant ψ and R≡0 this flux-form expresses the conservation of mass; see Smolarkiewicz et al. (2014), for further insights and discussion.

Like SL schemes of the IFS, and many other NWP models, the FTFV schemes are inherently two-time-level. Consequently, in order to achieve second-order accuracy in time their derivation (see Smolarkiewicz (2006) for an overview) converts temporal derivatives to spatial derivatives by exploiting the analytic dependencies of temporal and spatial partial derivatives in the governing equation, a procedure known as Cauchy-Kowalewski; cf. section 19 in Toro (2010). In contrast to MOL schemes, the FTFV schemes can be viewed as approximations to volume integrals of the governing PDEs in a 4D continuum. Unsurprisingly, the semi-implicit FTFV integrators for fluid PDEs can be written in a numerically congruent form to SL schemes. For example, assuming trapezoidal time integration of the rhs forcings R, leads to a single template, $\psi^{n+1} = A(\psi^n + 0.5\,\delta t R^n) + 0.5\,\delta t R^{n+1}$, common for trajectory wise SISL and finite-volume wise FTFV integration schemes (Smolarkiewicz and Margolin, 1993; Smolarkiewicz et al. , 2014). Moreover, in the SISL of the IFS only linearized terms R^{n+1} are treated implicitly, whereas FTFV schemes typically include the full non-linearity of R^{n+1}. For a semi-Lagrangian scheme, the operator A refers to mapping its argument, $\psi' \equiv \psi^n + 0.5\,\delta t R^n$, to the departure point of the trajectory, $(\mathbf{x}_o, t_o)$, in the 4D space; whereas for the FTFV integration it represents the solution of the homogeneous finite-volume advection of the argument, $A(\psi') \approx \psi' - (\delta t/\rho)\boldsymbol{\nabla}.\rho\mathbf{v}\psi'$. The numerical algorithm may be customized to adapt the explicit part of the scheme for a better representation of much shorter time-scales, e.g. associated with the disparity between the timescales of the fluid flow and the much shorter timescales associated with phase-change processes and precipitation (Grabowski and Smolarkiewicz, 2001).

The algorithmic similarity of the two-time-level SISL and FTFV integrators is practical, convenient and enabling. In particular, it empowers a class of non-oscillatory and sign preserving FTFV advection operators A (with different levels of accuracy, complexity and computational expense) that benefit computational stability and physical realisability of numerical results via implicit large eddy simulation (ILES; Smolarkiewicz and Margolin, 2007), while implying that the remaining part of the solution procedure --- i.e., formulating and solving a suitable elliptic BVP problem --- can be the same for, and thus shared by, the SISL and FTFV integrators. Furthermore, as the FTFV integrators enable both structured grid and unstructured mesh realisations (Smolarkiewicz et al., 2013; Szmelter et al., 2015) they can coexist and cooperate within the same model code with spectral transform based SISL integrators.

FVM is designed with the general aim to supplement the IFS with a complementary, highly scalable non-hydrostatic module enabling numerical procedures unavailable in SISL spectral models and with minimal disruption to the highly optimised parallel code. Because FVM operates at the nodes of the IFS grid, it seamlessly inherits the equal-area domain decomposition of the IFS, with multiple layers of parallelism hybridising MPI tasks and OpenMP threads; cf. sections 5.1 and 5.2. On the other hand, because FVM operates on local stencils, it can replace non-local communication and computation patterns with nearest-neighbour equivalents, providing an effective tool for assessing the utility of innovations from outside the realm of spectral methods in the context of real weather. The degree to which the FVM may offer a supplement or an alternative to specific elements of the IFS can vary from simple advection of selected fields in a conservative non-oscillatory finite-volume fashion, to replacing the whole dynamical core with a cloud-resolving model. The local stencil and the SI integrators of the FVM supplied with a general robust elliptic solver enable conservative, nonoscillatory integrations with a level of implicitness unavailable in IFS. In particular, FVM admits arbitrarily steep orography (section

5.6). However, unlike the SISL IFS, the conservative advection in FVM needs to observe stability criteria of explicit schemes, but this may be leveraged by continuing to use the dynamical fields advected by the SL scheme of IFS.

5.5 Transport of species

One of the greatest strengths of standard SL advection schemes (where interpolation coefficients do not depend on the advected field), along with the ability to integrate stably and accurately using long time steps, is the efficiency in multi-tracer transport applications such as environmental, atmospheric composition forecasts. In such applications, the cost of tracer advection reduces to the cost of interpolating each tracer field to the same departure point field, computed only once per time step (although all fields still need to be individually remapped to that departure point). However what is gained in computational efficiency is somewhat lost in mass conservation.

Global mass conservation errors with respect to individual tracers are generally larger than the corresponding errors for the air mass and depend on tracer features such as the "roughness", background values and the atmospheric region over which the tracer is spread. For example, experiments using the humidity field as a passive tracer show global mass conservation errors in a 10-day simulation approximately equal to 1%. For the less smooth cloud fields this increases to approximately 5%. Despite this, for real forecasts with moist physics processes at current hydrostatic resolutions, lack of mass conservation is of secondary importance for NWP timescales as the uncertainty in physical parameterization tendencies is typically much larger than the mass conservation error per time step. However, the cumulative effect in long time integrations or the local effects in non-hydrostatic cloud resolving simulations are not negligible. To address the local effects, in particular in convection permitting simulations, COMAD interpolation (Malardel and Ricard, 2015) has been developed, which improves considerably local mass conservation of moist species and additionally improves precipitation skill scores without any significant increase of the computational cost.

In the hydrostatic IFS atmospheric composition forecasts the approach currently followed is to use global mass fixers (Diamantakis and Flemming, 2014). The schemes that have been developed restore global mass conservation using a weighted approach which takes into account local transport errors as opposed to simpler algorithms which adjust gridpoint values at equal proportions defined by the ratio of the global mass before and after the advection step. This weighted approach is beneficial, particularly in cases where tracers have relatively large background values and strong gradients, such as CO_2 and CH_4, where the fixer is able to correct the transported field in areas of gradients, leaving the smooth background field mostly unchanged (Agusti-Panareda, 2014). Figure 13 shows the difference in the concentration of CO_2 due to the mass fixer for two simulations: one with a simple equal-proportion algorithm and another with the locally weighted approach of Bermejo and Conde, see Diamantakis and Flemming (2014) for details. We notice that the latter produces larger increments in areas of strong CO_2 emissions while the former unrealistically spreads the correction almost uniformly around the globe.

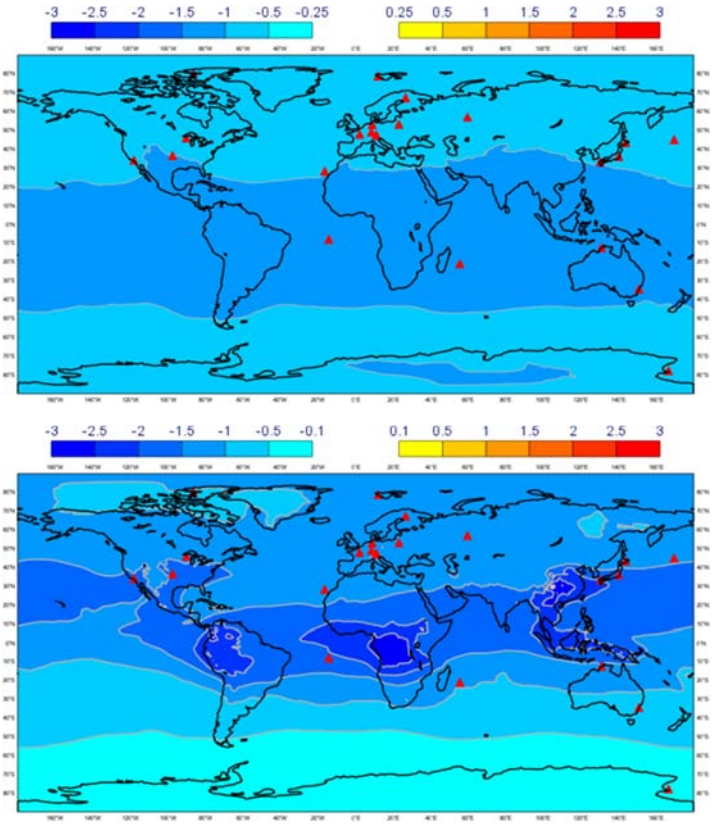

Figure 13: Total column average CO_2 [ppm] March 2012. Top: Difference of the simulation result without mass fixer minus the simulation result using a simple proportional mass fixer. Bottom: Difference of the simulation result without mass fixer minus the result using the more advanced locally weighted approach of Bermejo and Conde algorithm.

Global mass fixer algorithms do take into account local transport errors but they cannot guarantee local conservation and thus lack local (flow) consistency. This is because they are based on heuristic error criteria which do not conform to the continuity equation. Consistent local mass conservation is a desirable property for short lifetime, reactive species, and FVM (section 5.4) provides for the technical ability to implement an inherently local and global mass conserving scheme in the future.

5.6 Vertical discretization, orography treatment

The H-IFS employs a hybrid eta-p vertical coordinate (Simmons and Burridge, 1981) which is terrain-following near the surface and gradually converts into a pressure coordinate in the free atmosphere. The original finite difference (FD) vertical discretisation with a Lorenz grid staggering has been replaced by a higher-order, finite element (FE) scheme with co-located data arrangement (Untch and Hortal, 2004). The FE vertical discretisation improves forecasts especially in the stratosphere, where the solution dependence on physics parameterisations weakens and vertical mesh spacing coarsens. The same hybrid vertical coordinate based on hydrostatic pressure/mass has been retained in the NH-IFS, but implemented with a FD scheme (Bubnová et al., 1995). Notably, the NH-IFS maintains the separation of horizontal and vertical motions, thus facilitating essentially the same spectral transform solution procedure as for the H-IFS. A vertical FE scheme in the NH-IFS has been impeded by complications associated with the particular non-hydrostatic prognostic variables used for optimal stability (Bénard et

al., 2005), the presence of both integral and differential vertical operators (only integral operators occur in the H-IFS), and consistency constraints of the SI scheme. A successful implementation of a vertical FE scheme has recently been reported in the non-hydrostatic ALADIN limited-area model (Vivoda et al., 2012). It may become available in the NH-IFS within the next cycles for testing in global non-hydrostatic simulations.

Research into alternative numerical schemes and vertical discretisations is ongoing within the newly developed non-hydrostatic FVM that is intended to supplement the IFS. Here, the independent vertical coordinate on the computational mesh has been selected to be (geometric) height, in contrast to pressure in the H-IFS and NH-IFS. However, to provide flexibility with respect to the specification of the model levels, a generalised vertical coordinate is considered. The latter is embedded, via a consistent time-dependent curvilinear coordinate formulation, in the two-time-level SI integration scheme for the deep-atmosphere, non-hydrostatic, compressible Euler equations. Moreover, the formulation permits a departure from the spherical geometry towards the simulation on the geoid. The principally unstructured horizontal spatial discretisation is combined with a flux-form FD approach in the vertical direction, where a structured grid with co-located arrangement of variables is favoured (see Smolarkiewicz et al., 2015 and Kühnlein et al., 2015 for a comprehensive description).

Similar to NH-IFS, the SI scheme in the FVM has been designed to account for implicit treatment of fast acoustic and buoyant modes, along with implicit treatment of rotational modes. However, the direct spectral solution of the Helmholtz problem in the SI scheme of the H/NH-IFS demands the implicit part to be linearly separable on each model layer, which does not permit variations of the problem coefficients such as arising from orography. Thus, implicit treatment of orographic effects and the associated boundary conditions is fundamentally not possible with the existing direct spectral solver without breaking the separability of horizontal and vertical; see also (Bénard, 2014). In contrast, the SI scheme of the FVM considers problem coefficients of the elliptic operator that vary in all spatial directions and in time, facilitating an effective implicit incorporation of orography and static or dynamic adaptivity of the vertical coordinate. Moreover, vertical boundary conditions of the elliptic problem are also treated implicitly, rather than in a split manner between explicit and implicit parts of the time stepping scheme, thus providing better stability and consistency of the solution. A numerically consistent 3D implementation is achieved using a specifically tailored, preconditioned nonsymmetric generalized conjugate residual variational iterative elliptic solver (Smolarkiewicz et al., 2007).

An important aspect of the higher resolutions anticipated for future NWP is the concomitant increase in the realism of the orography that can be incorporated. On the one hand, this will eventually lead to a more comprehensive representation of orographically forced atmospheric processes, where current models show deficits. On the other hand, the increasingly complex (mean) orography is going to challenge existing numerical model formulations. The common approach in NWP models of using terrain-following vertical coordinates can potentially introduce significant truncation errors as a result of complex orography, and locally steep slopes of the terrain-following levels may trigger instability in some formulations. Nevertheless, although alternative techniques have been proposed to accommodate orography in NWP models (e.g. Steppeler et al., 2002; Szmelter et al., 2015), terrain-following coordinates remain attractive for global NWP.

Terrain-following coordinates represent a convenient approach to implement uniform high resolution along the orography together with a structured grid for efficient direct matrix inversions in the stiff vertical direction. Various techniques have been proposed to achieve more accurate computations with terrain-following coordinates/meshes and to push stability limits towards steeper slopes (e.g. Schär et al., 2002; Smolarkiewicz et al., 2007; Zängl 2012). This is in line with the FVM development. The FVM has been designed with the aim for a robust and efficient incorporation of complex high-resolution orography using a terrain-following, adaptive vertical coordinate. Various canonical problems of stratified flows past idealised orography in hydrostatic and non-hydrostatic regimes have shown good results, and testing with realistic orography as required in H/NH-IFS applications has started. So far, validation for various idealised mountain shapes confirm that maximum slopes of 80 degrees and beyond can be handled without stability problems. Figure 14 illustrates flows past three dimensional mountains with 1750m and 7000m height, respectively. The inviscid FVM solutions are similar to the results obtained using the ICON model (Zängl, 2012; Zängl et al., 2015), although not in amplitude and detail behind the mountain, which may be explained by the explicit vertical diffusion on momentum and temperature used in the latter.

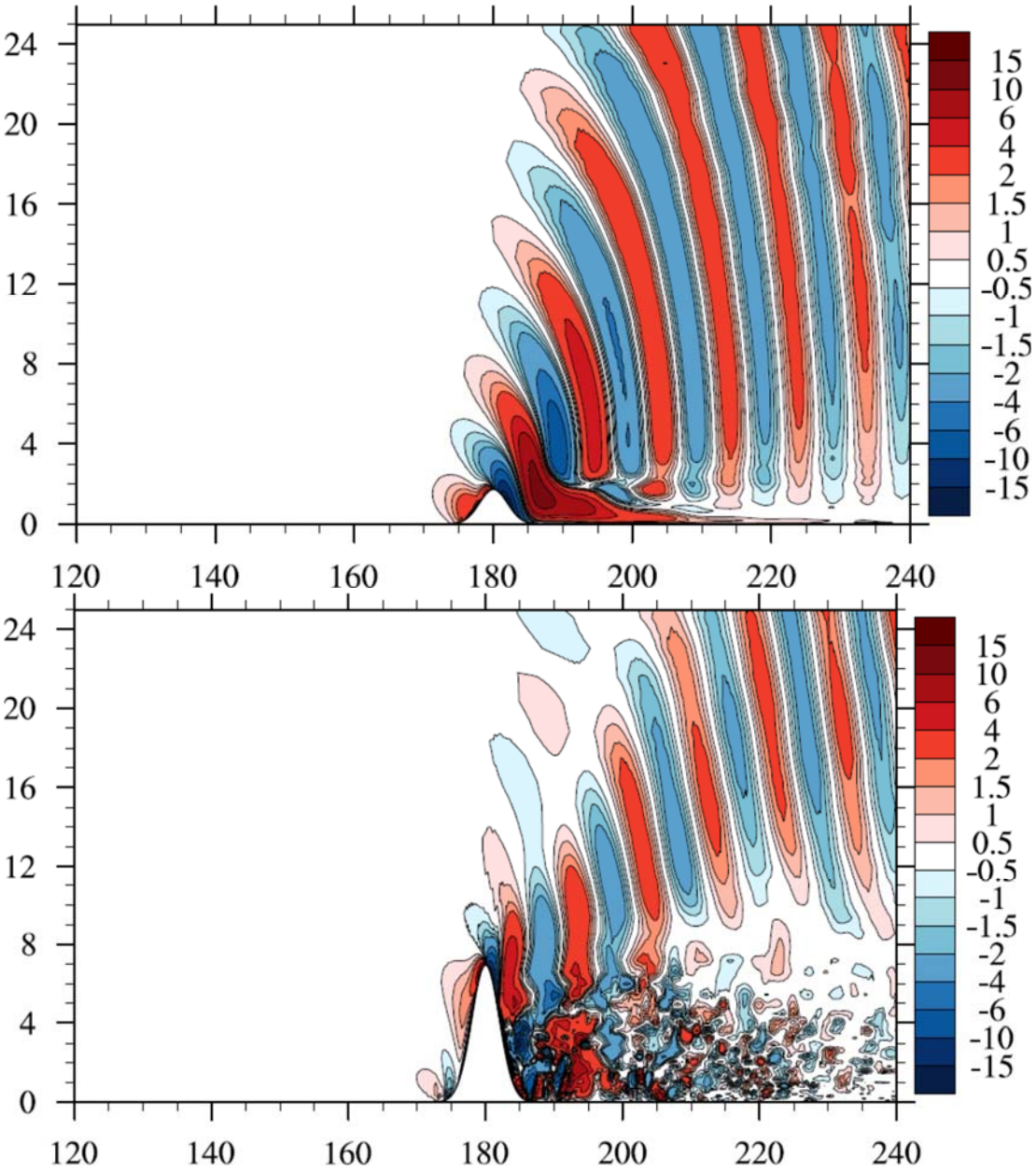

Figure 14: Vertical velocity (in m/s) from the simulation of stratified inviscid flow past an isolated Gaussian hill on a small planet with a radius of 40 km. The initialisation uses an isothermal stratification with 288 K, and a uniform flow of 20 m/s. The Gaussian hill has a half-width of 2 km, and hill heights of 1.75 km and 7 km are considered in the upper and lower panel, respectively, corresponding to maximum slopes of 36 and 71 degrees. Results are after 2 hours of simulation. In Zängl (2012) and Zängl et al. (2015) limited-area simulations with a similar setup are reported for the ICON model.

Idealised experiments with orography in a resting atmosphere conducted with the spectral SISL framework of the NH-IFS indicate stability limitations with orographic slopes beyond ~50 degrees. Moreover, accuracy issues in the form of significant spurious vertical velocities >5 m/s develop with 50 degree slopes (not shown).

Ongoing and future work investigates dedicated mesh generation procedures to attain improved terrain-following mesh quality in the presence of complex orography covering the globe (e.g. in the spirit of Thompson and Warsi, 1979; Schär et al., 2002; Klemp, 2011). Such procedures can be effectively applied in the static height coordinate framework underlying the FVM, but are less amenable to pressure-based coordinates.

Generally, the flexibility provided by the vertical coordinate formulation in the FVM will allow exploration of a broad range of vertical level specifications. As one example, the FVM framework will serve to revisit the current vertical mesh stretching based on a single global profile in the IFS. This will be done by introducing more general 3D mesh stretching that allows to better account for regional differences of the atmospheric structure, e.g. the height of the tropopause where a finer vertical spacing can be beneficial to better represent the UTLS region. Beyond static adaptivity of the model levels, dynamic adaptivity of the vertical coordinate will be considered and is already part of the design of FVM.

5.7 Physics-Dynamics coupling

The full set of equations of numerical models of the atmosphere results from averaging the equations valid at the continuum scale to the scale represented/resolved in the model. The latter may be determined by an explicitly specified filter scale or by the implicit filter scale given by the discretization. The dynamical core treats the interaction of the resolved motions including all adiabatic processes (see section 5.3). The remaining terms, i.e. terms which involve poorly resolved, sub-filter or subgrid-scale motions, as well as diabatic processes and water phase changes are essential to reproduce realistic atmospheric states in the model. These terms are computed separately in a suite of different process models, called the physical parametrisations (or the “physics”) that describe these diabatic processes, and that express the effect of subgrid scale motions on the resolved motions --- expressed in terms of the resolved variables and thus depending on them. Process splitting is a common and convenient solution to reduce the complexity of the otherwise complex implicit coupling between physics and dynamics and within the physics. The process splitting between dynamics and physics is based on a time and scale separation assumption between the ``resolved'' circulations and the parametrised subgrid or sub-filter scale processes. However, in the atmosphere there is no such separation between atmospheric scales.

As the resolution of the model increases, some subgrid processes currently targeted by parameterisation will become partially resolved. This is expected for the gravity wave drag parameterization at resolutions higher than 5 km, and for convection from resolutions less than 1 km. Radiation, water phase changes and the explicit interaction with the surface will never be directly resolved and must be linked to the given discretization of the prognostic variables. However, as the resolution increases, the nature of a parametrisation may change. For example, at resolutions O(1km) or less, the “turbulent” subgrid transport which is currently parametrised as a 1D vertical diffusion process requires the inclusion of 3D effects as is common in large eddy simulation (LES) models.

Physics/Dynamics coupling in the IFS

In the current IFS, "physics" comprises the radiation, turbulent diffusion, orographic and non-orographic gravity wave drag, convection and cloud microphysics, and at the surface boundary the land surface parameterization scheme, cf. section 5.10 for the latter.

The coupling strategy between the physics and the dynamics, and between the physical parametrisations themselves is mainly sequential: the result of the previous component forces the next within the large time steps possible due to the SL scheme.

The IFS parametrisations solve processes locally inside a grid box or along a vertical column neglecting horizontal exchanges between grid boxes. This concept is convenient for parallelization and appropriate as long as the processes modelled are truly subgrid scale. The physics computations are done on the same grid, and the same terrain-following hybrid model levels as the dynamics, except for radiation which is done on a coarser grid and with a different time step to reduce the cost of the radiative transport calculations. Consistent with the trajectory integral in section 5.4 the physics tendencies of the slower parametrised processes (radiation, convection) are computed at the departure point of the trajectory. The final "slow" tendencies are then estimated along the SL trajectory by the average between the tendency computed at the beginning and at the end of the trajectory. The fast processes (large scale condensation and precipitation, boundary layer processes) are treated as an instantaneous adjustment of the atmospheric state at the arrival point of the trajectory.

At each time step, the physics are forced by an explicit guess computed by the dynamical core. For example, an adiabatic cooling due to vertical advection in the dynamics generates supersaturation which is handled later in the physics. In addition, each parametrisation or group of parametrizations is solved sequentially and independently with its own implicit and/or iterative numerical solvers.

The physics finally provides tendencies for temperature, zonal and meridional wind, specific humidity, specific cloud, ice, rain, snow content and cloud cover to the physics/dynamics interface. There is no tendency computed for surface pressure, which means that the physics processes are computed at constant hydrostatic pressure (no mass redistribution, neither in the vertical nor the horizontal). Accordingly, the tendencies from parametrisations which produce tendencies for conservative variables like potential temperature or static energy are projected entirely on temperature (rather than temperature and pressure). The pressure and geopotential adjustment to diabatic heating is "handled" by the dynamical core when solving the implicit system, thus balancing the contributions of physics and dynamics.

The transition from a parametrised to a resolved process depends also on the effective resolution of the state of the atmosphere provided by the dynamical core to the physics parametrisations (see Lander and Hoskins, 1997 for a discussion). The recent change from a "linear" grid to a "cubic" grid in the IFS showed a clear improvement in the effective resolution as measured by improved predictive skill (not shown) together with a more realistic kinetic energy spectrum at a large range of higher wavenumbers (Figure 15). Moreover, due to the filter effect of the truncation much less horizontal diffusion is applied in the cubic grid simulations compared to the linear grid. In addition, highly non-linear processes like advection or surface-atmosphere interactions benefit from the finer gridpoint representation of the cubic grid. Notwithstanding, selected processes may be computed for efficiency on a coarser grid which could be explored in the future.

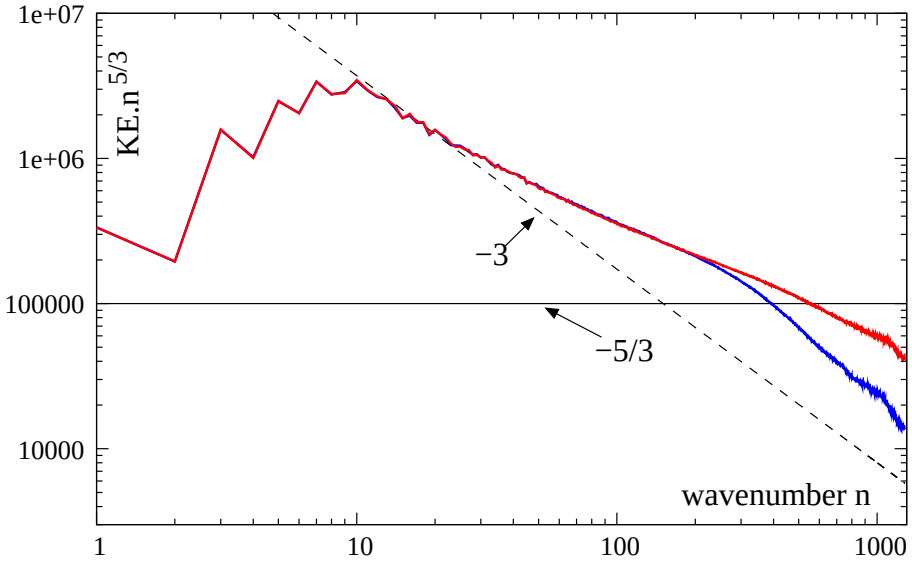

Figure 15: Globally integrated monthly average kinetic energy spectra as a function of total wavenumber, comparing the TCo1279 (red) and the TL1279 (blue). Spectra are scaled with the 5/3 power of the wavenumber so that solid and dashed black lines indicate a -3 and a -5/3 slope, respectively.

Evolution of Physics/Dynamics coupling

The current design of the IFS physics/dynamics coupling is typical of models with numerics allowing very long time steps. Models with a smaller time step often have a more explicit/parallel coupling where all the parametrisations are forced by the same atmospheric state, except for the processes which are still supposed to be much faster than the time step, e.g. the adjustment to supersaturation.

For an Eulerian flux form discretization such as in FVM with a time step limited by a CFL criterion (section 5.4), the current coupling strategy may not be appropriate anymore and the parametrisations themselves may have to evolve to be consistent with the numerics. Moreover, in search of exposing more parallelism in the model, it may be necessary to execute part or all of the physics in parallel rather than sequentially.

The analysis of the coupling of the current IFS parametrisations in the context of the NH-IFS confirmed the importance of the consistency between the physics and the equations solved (Malardel, 2011, 2014). Moreover, resolving non-hydrostatic effects happens at resolutions where convective motions are partly resolved and thus implies an inherent difficulty to accurately represent the flow associated with mesoscale convective systems. Efforts to develop scale-aware convective parametrisation started in the LAM community more than ten years ago (Gerard, 2007) and are now shared by a much larger community (Siebesma, 2015; Arakawa and Wu, 2013, 2014; Gustafson, 2013; Grell and Freitas, 2014). However, these efforts have not yet matured for global NWP. Recent results with variable resolution meshes (Müller, 2014) show the need for scale-aware physics across the computationally affordable but difficult range from 10km down to 1km grid distances. This is particularly important for the range of applications envisaged at ECMWF in the future, covering forecast ranges up to 1 year ahead with potentially multi-resolution setups.

Other problems in the design of the physics/dynamics coupling may arise when the design of the dynamical core evolves towards alternative discretizations. For example, the pointwise representation

of prognostic variables in higher order finite element methods used by some new dynamical cores (Table 1) is not fully consistent with the ``finite volume'' bulk representation assumed in most physical parametrisations (Lauritzen, 2014).

The validation of different physics/dynamics coupling strategies is difficult due to the block code structure adopted in many models. The best coupling solution may be different for different processes. Moreover, for efficiency reasons models using relatively small timesteps in the dynamics adopt much larger timesteps for different processes in the physics (Zängl et al., 2015). The impact of this is unclear. Thus in order not to limit the scientific choices, the concept of flexibility must be extended beyond the two main blocks of physics and dynamics that currently exist in IFS (with the notable exception of radiation which is already computed on a coarser grid).

The development of more local operators discussed in section 5.1 will facilitate new scientific developments in the physics and its coupling with the dynamics. In particular, nearest-neighbour communication and local computation of horizontal derivatives opens up the possibility to implement efficient 3D turbulence and possibly other 3D effects such as radiation and cloud interactions. Moreover, hierarchical grid structures will support multi-scale interfaces for different physical processes.

5.8 Tangent linear and adjoint model

The use at ECMWF of the 4D-Var algorithm for data assimilation makes it necessary to develop and maintain a tangent linear and adjoint forecast model, both for the dynamical core part (Temperton, 1998) as well as for the physical parameterizations (Janiskova and Lopez, 2012). For the current spectral transform model, we maintain tangent linear and adjoint versions of most options operationally used within the non-linear model. The development of this code, especially the SL scheme, took substantial effort amounting to several person years. The use of automatic adjoint code generators was considered but this approach has always been found to produce sub-optimal code, both in terms of efficiency and practicality. In general, the tangent linear and corresponding adjoint code used is the formal linearisation of the non-linear scheme but there are notable exceptions, e.g. the monotonic interpolation used within the SL scheme cannot be used, because it is not differentiable.

With a 4D-Var assimilation system, any alternative dynamical core option requires the development of tangent linear and adjoint versions, i.e. for the FVM option. This requires additional expertise in developing accurate tangent linear and adjoint formulations, which in return requires to commit additional resources in the future to the effort of developing this code. Moreover, since the tangent linear model does not necessarily share the same stability criteria as the non-linear model, increasing horizontal and vertical resolutions could lead to model instabilities and require careful evaluation.

The use of the tangent linear model within the incremental 4D-Var system and within other applications (e.g. the singular vector generation for ENS and adjoint based observation diagnostics) is always at substantially lower resolution than the resolution in the non-linear model. This is an efficiency compromise in order to make the system affordable given the operational constraints. The operational time constraint is more severe than for the non-linear model. Instead of the need to run at a speed of 240 forecast days/day in the non-linear model, the equivalent of 5-8 times that speed, i.e. 1200-2000 forecast days/day are needed to fit 4D-Var into the operational schedule. There are efforts to improve this constraint with time-parallel techniques that split a given window into independent chunks of work

(OOPS project). However, the requirement of being able to run fast at a modest model resolution emphasizes the strong scaling properties of the model over the weak scaling properties.

5.9 Uncertainty quantification

It is recognised that considering comprehensively all uncertainties related to the simulation and the initial data is essential for weather forecasting. Sources of uncertainty are broadly categorised in (a) *aleatory*, relating to inherent (random) variation that may be captured given sufficient statistical samples and characterized by a probability density distribution, and (b) *epistemic*, uncertainty due to lack of knowledge (Rider et al., 2010; Roy and Oberkampf, 2011). Sources of uncertainty in weather and climate simulations may be a mixture of the two, including insufficient intrinsic variability of the model, uncertainty arising from model parameters, numerical uncertainties, initial state and boundary condition uncertainty, lack of resolution, insufficient model complexity and lack of realism. Aleatory uncertainty is also termed irreducible or unavoidable and is addressed today by using ensembles of analyses and simulations and explicitly including stochasticity. Epistemic uncertainty however, is reducible or may be eliminated for selected processes. The topics addressed here aim to quantify and reduce epistemic uncertainty.

In preparation for extreme scale HPC systems, additional sources of uncertainty or even disruption of the simulation will arise from the hardware. In order to provide application resilience against such hardware related uncertainty, it is important to understand and quantify the uncertainties related to every individual component of the complex NWP workflow. This is formally addressed in VVUQ (verification, validation and uncertainty quantification) techniques for complex scientific simulations (Rider et al., 2010; Roy and Oberkampf, 2011). Especially with large HPC systems, the infrastructure developed at ECMWF, e.g. Atlas, facilitates a better validation hierarchy from simple to complex configurations.

One example of the multi-scale complexity of NWP workflows is the coupling of physics and dynamics. The large sensitivity of the results to the physics/dynamics coupling strategy shows the importance of the problem. Introducing probability distributions or a stochastic component in the physics/dynamics coupling mixes aleatory and epistemic uncertainty, but may be an effective technique to represent uncertainty in the thermodynamic vertical profiles on which the parameterizations act. The first Physics/Dynamics Coupling workshop (PDC2014) reviewed the ongoing development and testing of sensitivity with respect to physics/dynamics coupling in the context of the many emerging new dynamical cores (Gross et al., 2015). Intercomparison and reproducibility beyond typical dynamical core validation clearly represents a problem. Single-column model testing typically used in parameterization development is insufficient to represent the two-way coupling between physics and dynamics. In response, idealized tests specifically designed to evaluate the skill of the coupling are sought together with a community effort to provide a set of simplified parametrisations to further the understanding of physics/dynamics feedbacks and to facilitate complex model intercomparison.

5.10 Earth-System Coupling

The atmospheric prognostic variables in IFS are temperature, divergence and vorticity, humidity, surface pressure, cloud and precipitation variables as well as the ozone mass mixing ratio. At the sea surface, the IFS has a two-way coupling to the WAve Model (WAM, Hasselmann et al., 1988) and, in ensemble mode, IFS is coupled to the Nucleus for European Modelling of the Ocean (NEMO, Madec, 2008) as shown in Figure 16. The coupling with WAM is performed every time step whereby WAM is forced by IFS 10-metre wind speeds while the IFS is forced by surface roughness over sea. The most important prognostic variable provided by NEMO is the sea-surface temperature (SST) but also information from the ocean surface current is passed to the atmosphere. In the future, time evolving sea-ice information from a dynamic sea-ice model will be passed from NEMO to the atmosphere. NEMO, on the other hand, receives radiation, evaporation-precipitation and wind information (via the WAM model) to drive the ocean circulation. Over land surfaces, the parameterization scheme HTESSEL (Balsamo et al., 2011) is an integral part of the physical parameterizations and is called every time step affecting heat and moisture fluxes. HTESSEL can also be run-offline using atmospheric forcing from reanalyses.

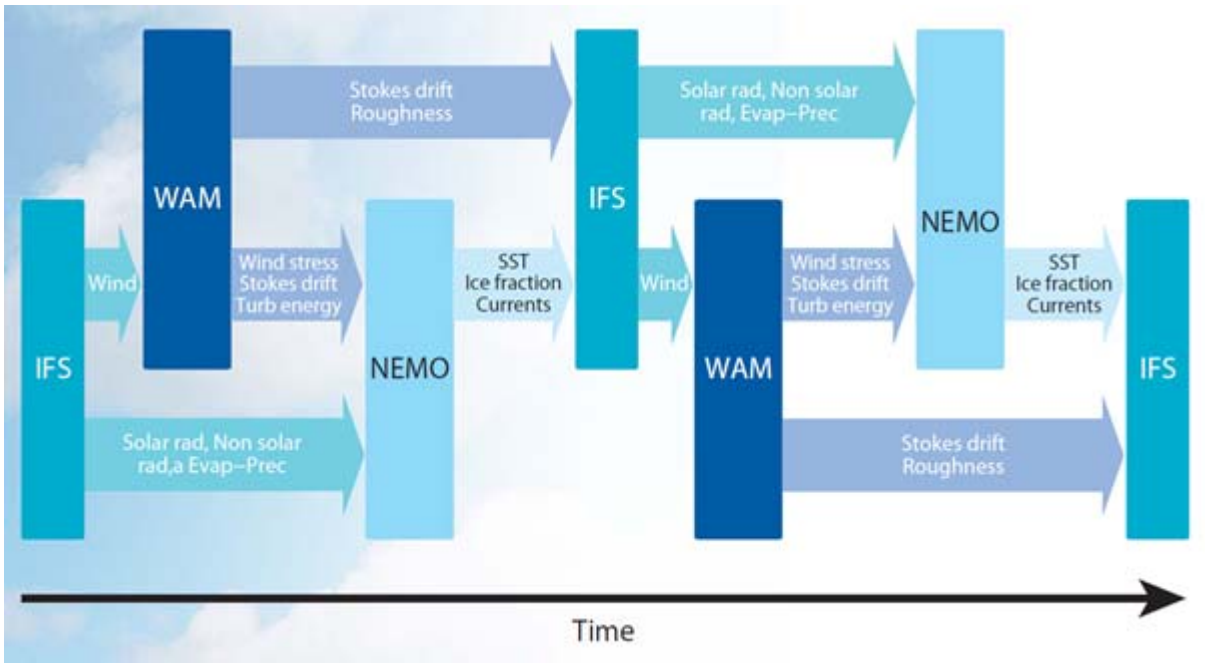

Figure 16: Coupling sequence and exchange of physical quantities between IFS, WAM and NEMO in the present model configuration (Mogensen 2014).

The additional computational cost of the coupled components is currently negligible compared to the atmospheric part of the high-resolution forecast but similar in cost for the present (lower resolution atmosphere) ensemble predictions, and the scalability of the coupled system is only marginally affected by coupling frequency. Since NEMO is part of the IFS executable there is no fundamental technical difference between calling the atmospheric physics (including HTESSEL) and the ocean model, apart from the re-gridding of model fields at every coupling step.

In the future, the coupling to Earth system components such as atmospheric composition, ocean, waves, sea-ice and land surfaces will become an even more integral part of the IFS, which has implications on

the model infrastructure and its initialization. Many technical developments that are currently being performed to advance code flexibility and efficiency (see section 5.1) will support this integration. The strategic goal is to have a fully coupled atmosphere – ocean – sea-ice model and to add numerous aerosol and trace gas species to the current set of prognostic variables, while maintaining the required NWP efficiency.

The integration of atmospheric composition adds a significant number of prognostic variables (in gridpoint space). Moreover, future enhancements of cloud microphysics schemes towards higher-moment representations of particle size spectra require a further increase in prognostic variables. If the set of chemical process parameterizations that is currently maintained in the Copernicus Atmospheric Monitoring Service (CAMS) system is activated in the IFS, the number of prognostic variables will increase from O(10) to O(100). To control computational cost, this may require running selected parameterizations or tracer advection on coarser grids and with fewer time steps than the physical part of the model. The technical developments described in section 5.1 and in particular the flexible data structures and domain decomposition facilitate such application.

In climate community models such as EC-Earth, both atmospheric composition and surface coupling is performed through external couplers (e.g. OASIS) that perform the synchronization and interpolation of model fields to be exchanged between individual model components. While this approach simplifies swapping model elements it reduces computational efficiency at present. However, future couplers (e.g. OASIS-MCT) are expected to employ highly parallelized data exchanges and be closely tied to parallel I/O servers to overcome this limitation. In order to assure continued flexibility in the configuration and coupling choices made by the EC-Earth community or at ECMWF, the envisaged code flexibility becomes even more important. In particular, IFS developments focus on coupled data assimilation to assure accurate initial conditions. However, the coupling choices for assimilation may differ from the needs in climate modelling, e.g. due to the different temporal frequency of ocean and atmospheric observations, compared to the daily coupling frequency in climate modelling.

The framework that is currently developed for the atmosphere could help coupling with ocean (including waves) and sea-ice models. Ideally, these would make use of the same data structures in the discretization together with efficient means for exchanging information on different meshes. For example, there is a need to improve load balancing that can strongly vary with seasonal sea-ice coverage and is affected by mesh-refinement near coasts, inland seas, and at the edge of sea-ice. Atlas data structures may be explored for this purpose.

However, a deviation from the NEMO framework and developing an independent ocean/sea-ice model with novel data structures is not considered. Similarly, this applies to the preparation of ocean initial conditions for the initialization of the coupled model, thus relying on the NEMO community to provide a scalable ocean model.

6 Conclusions

This paper has provided a comprehensive overview of the essential components for successful global medium-range forecasting and the assets of IFS in particular. The competitiveness of the existing hydrostatic IFS model has been illustrated, while at the same time stressing the difficulty for new generation non-hydrostatic models to match or exceed this computational performance as would be

required for resolutions resolving non-hydrostatic scales. The task to prepare the IFS for emerging heterogeneous hardware is demanding, since potentially communication and computational patterns need to change, while at the same time increasing code portability for fundamentally different hardware.

The substantial challenges ahead, to prepare the IFS for the anticipated applications at ECMWF in the future, will be met by following a hybrid strategy, augmenting the existing highly efficient hydrostatic dynamical core with the required flexibility, novel numerical algorithms and procedures to support ECMWF's strategy for the next decades. For this we have provided a roadmap, detailing the progress for different development topics, which split into technical priorities to introduce the required code flexibility and assure future portability as well as adaptation to emerging energy-efficient hardware, and into scientific priorities. These are to combine the strengths of the newly developed, principally autonomous non-hydrostatic finite-volume module FVM with the hydrostatic semi-Lagrangian spectral transform options of the IFS, to review the vertical discretization, and to carefully address physics-dynamics as well as Earth-system component coupling.

The progress to date includes the flexibility to explore different horizontal discretizations, the addition of a new, powerful 3D solver for elliptic problems arising from the implicit discretization of the non-hydrostatic system, an option for inherently conserving, monotone, multi-tracer transport, and developments towards a flexible vertical coordinate formulation.

One may ask what is new compared to several decades ago. Beyond accuracy and stability, a new performance metric of equal importance has emerged that measures energy-to-solution in addition to time-to-solution. Energy-to-solution may present the biggest challenge for running global NWP models at km-scale resolutions in the future, and the application of novel numerical methods as discussed in this paper that require less communication and associated data movement will address this. However, what is really new is that the next decades will see adaptive numerical techniques and tools mature --- implemented with unprecedented efficiency --- that are essential to answer critical questions in NWP with direct relevance to ECMWF's operational applications.

The necessary NWP advances will be addressed on the one hand by continuously improving accuracy and removing gaps in our knowledge. On the other hand they will also be addressed by creating the efficiency required to establish the reliability of the forecasts, and by providing the reliability and application resilience required in a computing environment that by itself may be subject to disturbances or partial hardware failure. Both should be achieved through adaptive and self-correcting algorithms to advance reliable forecast skill beyond today's limits.

Acknowledgements

We would like to thank Glenn Carver, Irina Sandu, Elias Holm, Anton Beljaars, Roberto Buizza, Alan Thorpe, Erland Källén and Rupert Klein for their comments and suggestions on previous versions of the manuscript. Furthermore, we would like to acknowledge the helpful discussions with Pierre Bénard and Fabrice Voitus.

References

Arakawa, A., C. S. Konor, 2009: Unification of the anelastic and quasi-hydrostatic systems of equations, Mon. Weather. Rev., 137, 710-726.

Arakawa, A. and C.-M. Wu, 2013: A unified representation of deep convection in numerical modelling of the atmosphere. Part I. J. Atmos. Sci., 70, 1977-1992.

Balsamo, G., S. Boussetta, E. Dutra. A. Beljaars, P. Viterbo, and B. Van Den Hurk, 2011: Evolution of land-surface processes in the IFS. ECMWF Newsletter No. 127, 17-22.

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 Comput., 21, 1621–1638.

Beljaars, A., P. Bechtold, M. Kohler, J.-J. Morcrette, A. Tompkins, P. Viterbo, and N. Wedi, 2004: The numeric of physical parametrization. Proc ECMWF Workshop on recent developments in numerical methods for atmosphere and ocean modelling, European Centre for Medium-Range Weather Forecasts, Reading, U.K., September 2004.

Bénard, P., J. Vivoda, J. Masek, P. Smolíková, K. Yessad, R. Brozkova and J.- F. Geleyn, 2010: Dynamical kernel of the Aladin-NH spectral limited-area model : Revised formulation and sensitivity experiments. Quart. J. of the Royal Met. Soc., 136, 155-169.

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

Bénard, P., 2014: A non-hydrostatic SI dynamical core: current-state, limitations and perspectives. ECMWF Seminar Proceedings. Seminar on Recent Developments in Numerical Methods for Atmosphere and Ocean Modelling, Reading, UK.

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. Wea. Rev., 123, 515–535.

Cossette, J-F, P.K. Smolarkiewicz and P. Charbonneau, 2014: The Monge-Ampère trajectory correction for semi-Lagrangian schemes, J. Comp. Phys., 27, 208-229.

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

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

Davies, T., A. Staniforth, N. Wood, J. Thuburn, 2003: Validity of anelastic and other equation sets as inferred from normal-mode analysis, Q. J. Roy. Meteorol. Soc., 129, 2761-2775.

Dennis, J. M., Edwards, J., Evans, K. J., Guba, O., Lauritzen, P. H., Mirin, A. A., St-Cyr, A., Taylor, M. A., and P.H. Worley, 2012: CAM-SE: a scalable spectral element dynamical core for the Community Atmosphere Model, Int. J. High Perform. Comput. Appl., 26, 74–89.

Diamantakis, M. and J. Flemming (2014): Global mass fixer algorithms for conservative tracer transport in the ECMWF model. GMD 7, 965-979.

Dörnbrack, A., J.D. Doyle, T.P. Lane, R.D. Sharman, P.K. Smolarkiewicz, 2005: On physical realisability and uncertainty of numerical solutions, Atmos. Sci. Lett., 6, 118–122.

Dubos, T., F. Voitus, 2014: A semihydrostatic theory of gravity-dominated compressible flow, J. Atmos. Sci., 71, 4621-4638.

Durran, D. R., 1989: Improving the anelastic approximation, J. Atmos. Sci., 46, 1453-1461.

Durran, D.R., 2008: A physically motivated approach for filtering acoustic waves from the equations governing compressible stratified flow, J. Fluid Mech., 601, 365-379.

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. Rep. 2, Institut for Teoretisk Meteorologi, University of Copenhagen, 37 pp.

Fuhrer, O., C. Osuna, X. Lapillonne, T. Gysi, B. Cumming, B. Mauro, A. Arteaga, T.C. Schlthess, 2014: Towards a performance portable, architecture agnostic implementation strategy for weather and climate models, Supercomputing frontiers and innovations, 1, p. 45-62.

Gassmann, 2013: A global hexagonal C-grid non-hydrostatic dynamical core (ICON-IAP) designed for energetic consistency, Q.J.R. Meteorol. Soc., 139, 152-175.

Gerard, L., 2007: An integrated package for subgrid convection, clouds and precipitation compatible with meso-gamma scales, Q.J.R. Meteorol. Soc., 133, 711-730.

Giraldo, F. X., Kelly, J. F., and Constantinescu, E. M., 2013: Implicit–explicit formulations for a 3D Nonhydrostatic Unified Model of the Atmosphere (NUMA), SIAM J. Sci. Comput., 35, B1162–B1194.

Grabowski, W. W. and P.K. Smolarkiewicz, 2002: A Multiscale Anelastic Model for Meteorological Research, Mon. Wea. Rev., 130, 939–956.

Grell, G. A., and S. R. Freitas, 2014: A scale and aerosol aware stochastic convective parametrization for weather and air quality modelling. Atmos. Chem. Phys., 14, 5233-5250.

Gross, M., S. Malardel, Ch. Jablonowski and N. Wood, 2015: Bridging the (knowledge) gap between physics and dynamic, BAMS, in review.

Gustafson, W. I., P. –L. Ma, H. Xiao, B. Singh, P. J. Rasch, and J. D. Fast, 2013: The Separate Physics and Dynamics Experiment (SPADE)framework for determining resolution awareness: A case study of microphysics. J. Geophys. Res. (Atmos.), 118, 9258-9276.

Hasselmann, S., K. Hasselmann, P. A. E. M. Janssen, E. Bauer, G. J. Komen, L. Bertotti, P. Lionello, A. Guillaume, V. C. Cardone, J. A. Greenwood, 1988: The WAM model - A third generation ocean wave prediction model. J. Phys. Ocean., 18, 1775-1810.

Hortal, M. and A.J. Simmons, 1991: Use of reduced Gaussian grids in spectral models. Mon. Wea. Rev., 119, 1057–1074.

Horten, G. and S. Vandewalle, 1995: A space-time multigrid method for parabolic partial differential equations, SIAM J. Sci. Comput., 16(4), 848-864.

Hundsdorfer, W., J.G. Verver, 2003: Numerical Solution of Time-Dependent Advection-Diffusion-Reaction Equations, Springer, 483 pp.

Ishida, J, C. Muroi, K. Kawano, and Y. Kitamura, 2010: Development of a New Nonhydrostatic Model ASUCA at JMA. CAS/JSC WGNE Research Activities in Atmospheric and Oceanic Modeling.

Jameson, A., W. Schmidt, E. Turkel, 1981: Numerical Solutions of the Euler Equations by Finite Volume Methods using Runge-Kutta Time-Stepping Schemes, AIAA Paper 81-1259 , AIAA 14th Fluid and Plasma Dynamics Conference, Palo Alto, California.

Janjic, Z., and R.L. Gall, 2012: Scientific documentation of the NCEP nonhydrostatic multiscale model on the B grid (NMMB). Part 1 Dynamics. NCAR Technical Note NCAR/TN-489+STR, DOI: 10.5065/D6WH2MZX.

Janiskova, M., P. Lopez, 2012: Linearized physics for data assimilation at ECMWF, in *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. II)*, Park, S. K., Xu, L. (Eds.) chapter 11, 251-286.

Kaas, E., 2008: A simple and efficient locally mass-conserving semi-Lagrangian transport scheme. Tellus 60A, 305-320.

Klein, R., 2009: Asymptotics, structure, and integration of sound-proof atmospheric flow equations, Theor. Comp. Fluid. Dyn., 3, 161-195.

Klemp, J. B., 2011: A terrain-following coordinate with smoothed coordinate surface. Mon. Weather. Rev., 139, 2163-2169.

Kodama, C., M. Terai, A. T. Noda, Y. Yamada, M. Satoh, T. Seiki, S.-I. Iga, H. Yashiro, H. Tomita, K. Minami, 2014: Scalable rank-mapping algorithm for an icosahedral grid system on the massive parallel computer with a 3-D torus network, 40(8), 362-373.

Kuell, V., A. Gassmann and A. Bott, 2007: Towards a new hybrid cumulus parametrization scheme for use in non-hydrostatic weather prediction models. Q.J.R. Meteorol. Soc. 133, 479-490.

Kurowski, M. J., W. W. Grabowski, P. K. Smolarkiewicz, 2015: Anelastic and compressible simulation of moist dynamics at planetary scales, J. Atmos. Sci., 72, 3975-3995.

Kühnlein, C., et al.: An all-scale finite-volume solver for global atmospheric dynamics with a flexible terrain-following coordinate and steep orography capability. (in preparation for Geosci. Model Dev.).

Lander, J., and B. J. Hoskins, 1997: Believable scales and parametrizations in a spectral transform model. Mon. Wea. Rev., 125, 292-303.

Lauritzen, P. H., 2014: Physics-Dynamics coupling with Galerkin methods: equal area physics grid, PDC14, CIRESE, Ensenada, Mexico, Dec. 2014.

Lauritzen, P. H., R. D. Nair, and P. A. Ullrich, 2010: A conservative semi-Lagrangian multi tracer transport scheme (CSLAM) on the cubed-sphere grid, J. Comput. Phys., 229, 1401-1424.

Lee, J.-L. and A.E. MacDonald, 2009: A finite-volume icosahedral shallow water model on local coordinate, Mon. Wea. Rev. 137, 1422–1437.

Leopardi, P., 2006: A Partitioning of the Unit Sphere of Equal Area and Small Diameter, Electronic Transactions on Numerical Analysis, 25, 309--327.

Lipps, F. B., R. S. Hemler, 1982: A scale analysis of deep convection and some related numerical calculations, J. Atmos. Sci., 39, 2192-2210.

Knoll, D.A., L. Chacon, L.G. Margolin, V.A. Mousseau, 2003: On balanced approximations for time integration of multiple time scale systems, J. Comput. Phys., 185, 583–611.

Madec, G., 2008: NEMO – the OPA9 ocean engine: Note du Pole de Modelisation, Institut Pierre-Simion Laplace, 1:100, available at: http://www.nemo-ocean.eu.

Malardel, S. and N.P. Wedi (2015): Hoes does subgrid scale parameterisation influence non-linear spectral energy fluxes in global NWP models ?, J. Geophys. Res, submitted.

Malardel, S. and D. Ricard (2015). An alternative cell-averaged departure point reconstruction for pointwise semi-lagrangian transport schemes. Q.J.R. Meteorol. Soc. To Appear. DOI: 10.1002/qj.2509.

Malardel, S., 2011: Physics/dynamics coupling. ECMWF Workshop on Non-hydrostatic Modelling, 8-10 November 2010, Reading, U. K., 67-78.

Malardel, S., 2014: Physics/dynamics coupling at very high resolution: permitted versus parametrized convection. ECMWF Seminar on Recent Developments in Numerical Methods for Atmosphere and Ocean Modelling, 2-5 September 2013, Reading, U. K., 83-98.

Marras, S., J. F. Kelly, M. Moragues, A. Müller, M. Kopera, M. Vázquez, F. X. Giraldo, G. Houzeaux, O. Jorba, 2015: A Review of Element-Based Galerkin Methods for Numerical Weather Prediction: Finite Elements, Spectral Elements, and Discontinuous Galerkin, *Archives of Computational Methods in Engineering,* to appear.

Michalakes, J., M. Govett, R. Benson, T. Black, H. Juang, A. Reinecke, B. Skamarock, 2015: AVEC Report: NGGPS Level-1 Benchmarks and software evaluation, submitted.

Mogensen, K., 2014: The coupled ocean-atmosphere model at ECMWF: overview and technical challenges. 16th ECMWF Workshop on High Performance Computing, available from http://www.ecmwf.int/en/workshop-high-performance-computing-meteorology.

Mozdzynski G., A new partitioning approach for ECMWF's integrated forecasting system (IFS), p 148-166 in Proceedings of the Twelfth ECMWF Workshop: Use of High Performance Computing in Meteorology, 30 October - 3 November, 2006, Reading, UK, World Scientific (2007) 273 pp.

Mozdzynski, G., M. Hamrud, N.P. Wedi (2015), A Partitioned Global Address Space implementation of the European Centre for Medium Range Weather Forecasts Integrated Forecasting System, Int. J. High Perform. Comput. Appl., 29(3), 261–273.

Müller, A., 2014: Does high order and dynamic adaptive refinement improve the efficiency of atmospheric simulation? , PDC14, CIRESE, Ensenada, Mexico, Dec. 2014.

Ogura, Y., N. A. Phillips, 1962: Scale analysis of deep and shallow convection in the atmosphere, J. Atmos. Sci., 19, 173-179.

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.

Park, H., S-Y Hong, H-B Cheong, and M-S Koo, 2013: A Double Fourier Series (DFS) Dynamical Core in a Global Atmospheric Model with Full Physics, Mon. Wea. Rev., 141, 3052—3061.

Park, S., 2014: Current Status and Update Plan for NWP and Typhoon Forecast Systems in KMA, Workshop on Numerical Prediction of Tropical Cyclone CWB, Taiwan, 20-22 May 2014, avail online.

Putman, W. L. and S.-J. Lin, 2007: Finite-volume transport on various cubed-sphere grids, J. of Comp. Phys., 227 (1), 55-78.

Qaddouri, A. and V. Lee, 2011: The Canadian Global Environmental Multiscale model on the Yin-Yang grid system, Q. J. R. Meteorol. Soc., 137, 1913 – 1926.

Rider, W. J., J.R. Kamm, G. Weirs, 2010: Verification, Validation, and uncertainty quantification workflow in CASL, SAND2010-234P.

Ritchie, H., 1988: Application of the semi-Lagrangian method to a spectral model of the shallow water equations. Mon. Wea. Rev., 116, 1587–159.

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

Roy, C. J., W. L. Oberkampf, 2011: A comprehensive framework for verification, validation, and uncertainty quantification in scientific computing, Comput. Methods Appl. Mech. Engrg., 200 2131–2144.

Satoh, M., H. Tomita, H. Yashiro, H. Miura, C. Kodama, T. Seiki, A. T. Noda, Y. Yamada, D. Goto, M. Sawada, T. Miyoshi, Y. Niwa, M. Hara, T. Ohno, S.-I. Iga, T. Arakawa, T. Inoue and H. Kubokawa, 2014: The Non-hydrostatic Icosahedral Atmospheric Model: description and development, Progress in Earth and Planetary Science, 1:18.

Schär, C., Leuenberger D., Fuhrer O., Lüthi D., Girard C., 2002: A new terrain-following vertical coordinate formulation for atmospheric prediction models. Mon. Weather. Rev., 130, 2459-2480.

Siebesma, A. P., 2015: First Workshop of the Grey Zone Project, GEWEX news, 25, 7-9.

Simmons, A. J., Burridge D. M., 1981: An energy and angular-momentum conserving vertical finite-difference scheme and hybrid vertical coordinates. Mon. Wea. Rev., 109, 758-766.

Skamarock, W.C., J.B. Klemp, 2008: A time-split nonhydrostatic atmospheric model for weather research and forecasting applications, J. Comput. Phys., 227, 3465–3485.

Skamarock, W.C, J. B. Klemp, M. G. Duda, L. D. Fowler, S.-H. Park, T. D. Ringler, 2012: A Multiscale Nonhydrostatic Atmospheric Model Using Centroidal Voronoi Tesselations and C-Grid Staggering, Mon. Wea. Rev., 140, 3090–3105.

Smolarkiewicz, P.K., Sharman R., Weil J., Perry S. G., Heist D., Bowker G., 2007: Building resolving large-eddy simulations and comparison with wind tunnel experiments, J. Comp. Phys., 227, 633-653.

Smolarkiewicz, P.K., J.A. Pudykiewicz, 1992: A class of semi-Lagrangian approximations for fluids, J. Atmos. Sci., 49, 2082–2096.

Smolarkiewicz, P.K., L.G. Margolin, 1993: On forward-in-time differencing for fluids: Extension to a curvilinear framework, Mon. Wea. Rev., 121, 1847-1859.

Smolarkiewicz, P.K., L.G. Margolin, 2007: Studies in geophysics. In Implicit Large Eddy Simulation: Computing Turbulent Fluid Dynamics, Chapter 14, Grinstein FF, Margolin L, Rider W (eds). Cambridge Academic Press: Cambridge, 413-438.

Smolarkiewicz, P.K., J. Szmelter, A.A. Wyszogrodzki, 2013: An unstructured-mesh atmospheric model for nonhydrostatic dynamics, J. Comput. Phys., 254, 184--199.

Smolarkiewicz, P.K., C. Kühnlein, N.P. Wedi, 2014: A consistent framework for discrete integrations of soundproof and compressible PDEs of atmospheric dynamics, J. Comput. Phys. 263, 185–205.

Smolarkiewicz, P.K., W. Deconinck, M. Hamrud, C. Kühnlein, G. Mozdzynski, J. Szmelter, N.P. Wedi, 2015: A finite-volume module for simulating global all-scale atmospheric flows, J. Comput. Phys., submitted.

Steppeler, J., Bitzer H., Minotte M., Bonaventura L., 2002: Nonhydrostatic atmospheric modeling using a z-coordinate representation. Mon. Wea. Rev., 130, 2143–2149.

Szmelter, J., Zhang Z., Smolarkiewicz P.K., 2015: An unstructured-mesh atmospheric model for nonhydrostatic dynamics: Towards optimal mesh resolution. J. Comp. Phys., 294, 363-381.

Szmelter, J., Z. Zhang, P.K. Smolarkiewicz, 2015: An unstructured-mesh atmospheric model for nonhydrostatic dynamics: towards optimal mesh resolution, J. Comput. Phys., 294, 363-381.

Szmelter, J., P.K. Smolarkiewicz, 2010: An edge-based unstructured mesh discretisation in geospherical framework, J. Comput. Phys., 229, 4980--4995.

Temperton, C., 1998: Tangent-linear and adjoint models, Seminar on Recent Developments in Numerical Methods for Atmospheric Modelling, 7-11 September 1998, Reading, UK.

Temperton, C. M. Hortal, A. Simmons, 2001: A two-time-level semi-Lagrangian global spectral model, Q.J. Roy. Meteorol. Soc., 127, 111--127.

Thompson, J. F., Warsi Z. U. A., Wayne Mastin C., 1982: Boundary-fitted coordinate systems for numerical solution of partial differential equations – a review, J. Comp. Phys., 47, 1-108.

Toro, E.E., 2010: Riemann Solvers and Numerical Methods for Fluid Dynamics, 3ed Edition, Springer, 724~pp.

Van Lent, J. 2015: Numerical integration on the sphere using an equal area mapping from a regular octahedron, in preparation.

Vivoda, J., Smolíková P., Simarro J., 2014: Vertical Finite Elements in the NH dynamical core. http://www.cnrm.meteo.fr/aladin/IMG/pdf/2013-02_VFE_Vivoda.pdf.

Voitus F., C. Kühnlein, P. Bénard, N. Wedi, 2015: Design of a constant-coefficient semi-implicit integration scheme for the unified system in a mass-based vertical coordinate, Q. J. Roy. Meteorol. Soc., in preparation.

Wedi, N.P. and P.K. Smolarkiewicz, 2006: Direct numerical simulation of the Plumb–McEwan laboratory analog of the QBO, J. Atmos. Sci., 63, 3326–3352.

Wedi, N.P., 2014: Increasing horizontal resolution in numerical weather prediction and climate simulations: illusion or panacea?, Phil Trans R Soc A, 372, 20130289.

Wedi, N. P., and P.K. Smolarkiewicz, 2009: A framework for testing global non-hydrostatic models, Q.J.R. Meteorol. Soc., 135 (639), 469–484.

Wedi, N. P., M. Hamrud, and G. Mozdzynski, 2013: A fast spherical harmonics transform for global NWP and climate models, Mon. Wea. Rev., 141, 3450-3461.

Wedi, N.P., K. Yessad and A. Untch, 2009: The non-hydrostatic global IFS/ARPEGE model: model formulation and testing, ECMWF Tech Memorandum 594, Reading, UK.

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

Wu, C.-M. and A. Arakawa, 2014: A unified representation of deep convection in numerical modelling of the atmosphere. Part II. J. Atmos. Sci., 71, 2089-2103.

Yessad, K. and N.P. Wedi, 2011: The hydrostatic and non-hydrostatic global model IFS/ARPEGE: deep-layer model formulation and testing, ECMWF Tech Memorandum 657, Reading, UK.

Zängl, G., D. Reinert, P. Rıpodas and M. Baldauf, 2015: The ICON (ICOsahedral Non-hydrostatic) modelling framework of DWD and MPI-M: Description of the non-hydrostatic dynamical core, Q.J.R. Meteorol. Soc., 141, 563-579.

Zängl, G., 2012: Extending the numerical stability limit of terrain-following coordinate models over steep slopes. Mon. Weather. Rev., 140, 3722–3733.

Zerroukat, M. and T. Allen, 2012: A three-dimensional monotone and conservative semi-Lagrangian scheme (SLICE-3D) for transport problems, Q.J.R. Meteorol. Soc., 138, 1640-1651.