Krylov solvers in the semi-implicit semi-Lagrangian AROME model to improve scalability of its future dynamical core

Thomas Burgot, Ludovic Auger, Pierre Benard

Krylov solvers in the semi-implicit semi-Lagrangian AROME model to improve scalability of its future dynamical core

Th. Burgot*, L. Auger, and P. Bénard
Météo-France, Toulouse, France
*Correspondence to: Thomas Burgot, CNRM/GMAP, 42 Avenue G. Coriolis, F-31057 Toulouse Cedex, France

To circumvent the scalability problem due to global communication involved in spectral transforms, a new version of the dynamical core where all calculations are performed in grid-point space is beeing built for AROME, the Météo-France's operational limited-area model. It is shown in an idealized but nevertheless realistic framework that despite this major change, keeping the other main characteristics of the model (constant-coefficient semi-implicit scheme, semi-Lagrangian transport scheme, A-grid, mass based coordinate, etc.) is possible. A Krylov solver is used to solve the implicit problem. A synthetic metric is then defined to compare the scalability of this method with the one of split-explicit schemes traditionally used in HEVI (Horizontally-Explicit/Vertically-Implicit) models. The latter are considered particularly scalable in the current parallelization paradigm and are used as a reference for scalability in this study. Using the solution given by the spectral model as a reference in terms of quality and required accuracy in an operational context, the chosen parameters of the Krylov solver are carefully tuned to maximize its convergence speed. Scalability measurements are then derived and it is shown that this new version is at least as scalable as the split-explicit schemes of the HEVI models, regardless of the future strategy considered for temporal and spatial resolutions: increasing spatial resolution while either maintaining a high Courant number or reducing it.

Key Words: scalability, Krylov solver, semi-implicit, dynamical core, NWP

Received . . .

Citation: . . .

1. Introduction

The quality of Numerical Weather Prediction (NWP) models has dramatically improved in recent decades (Bauer et al. 2015), partly due to the steady increase in spatial resolution, which itself was made possible by an increase of computing power. Operational limited-area models (LAM) commonly reach resolutions of about one kilometre or less in the horizontal and ten metres in the vertical. In this range of resolutions, a formulation of forecast models in fully compressible Euler equations is desirable since poorer governing systems (as Hydrostatic Primitive Equations) generate detrimental errors in parts of the domain where severe conditions occur.

In the last few years, supercomputers architecture has become more and more massively-parallel and this is likely to continue in the future. Parallelisation techniques have led to split the three-dimensional (3D) integration domain into many sub-domains, and the forecast is performed as a number of corresponding tasks that are handled in parallel. It is generally chosen to split the domain horizontally, thereby defining a tesselation of the geographical domain, whilst no split occurs along the vertical. By doing so, data

communications are required at each time step to exchange sub-domains information. Numerical methods that make the best use of parallelism are those that do communicate task data as little as possible with other tasks.

In order to cope with parallelization problems, some Meteorological Centres have chosen to use HEVI (Horizontally-Explicit/Vertically-Implicit) algorithms, whose implicit treatment only involves source terms responsible of the propagation of the fastest waves along the vertical, leading to the inversion of a one-dimensional (1D) problem along the vertical (see for example Satoh (2002) or Klemp et al. (2007)). Other source terms –which includes all terms involving discrete horizontal operators– are time-discretized in an explicit manner. The CFL constraint on the fast-wave part of horizontal motions thus imposes relatively small time steps. In the framework of the current massively-parallelized paradigm, these methods seem particularly well-adapted, and will be regarded in this study as a reference in terms of scalability.

Others Centres made the choice to use a semi-implicit (SI) treatment of all terms potentially responsible for the fastest waves propagation. This approach is often associated with a semi-Lagrangian (SL) algorithm for the transport part of the evolution. This combination (SI-SL) has made possible to considerably alleviate stability constraints, and thus to use large time steps, at the price of solving a 3D problem in space generated by the SI scheme. In the SI method, source terms are split into an appropriately chosen linear part and a so-called non-linear residual part. The system associated to linear terms allows a realistic propagation of fast-waves, and this part of the evolution is treated implicitly in time, while residuals are treated explicitly. When the coefficients of source terms in the linear system are horizontally-homogeneous and independent of time, the scheme is termed 'constant-coefficient SI scheme' (Bénard 2003).

In a 'constant-coefficient' SI scheme horizontal and vertical parts of the 3D-space problem to be solved are separated, making attractive a spectral-transform approach for the horizontal discretization, since direct methods may be used to solve the horizontal part of the problem. In the spectral-transform approach, (forth– and back–) transforms are carried out at each time step from the grid-point space, where non-linear (residual) terms are computed, to the spectral space, where several processes are performed: application of numerical diffusion, evaluation of discrete differential operators, and solution of the –separated– 3D implicit problem by a direct method. This is the strategy adopted so far for the AROME model, Météo-France's non-hydrostatic (NH) limited-area model (Seity et al. 2011).

However, a disadvantage of using time steps corresponding to large wave-CFL values is that the propagation of *all* fast waves –including gravity waves– is distorted. This implies namely that some distortion occurs in the response to orographic forcing. It may be argued that these wave-propagation errors may eventually exceed those generated by using a constant-coefficient approach instead of an approach retaining actual horizontal variations in linear terms, then leading to a non-separable 3D problem to be solved (e.g. Simmons et al. (1978)). Some studies even suggest that a fully-accurate representation of orographic effects can only be achieved for wind-CFL numbers below unity in SI-SL models (Pinty et al. 1995). Furthermore, spectral transforms require global communication of data among all computation nodes. Although spectral SI-SL schemes have shown an unchallenged efficiency and robustness, it is not sure whether this advantage will be maintained in the future, since costs of global communication could become prohibitive in operational contexts, finally raising an unsurmountable scalability problem. In current SI-SL applications, spectral transforms are therefore called as rarely as possible, by using very large time steps. These are carefully adjusted at each change in horizontal resolution in such a way to maintain the CFL number of fastest horizontal waves significantly greater than unity (about 14 currently for AROME).

Abandoning the spectral-transform SI-SL method for a SI-SL grid-point method with an iterative solver is the alternative strategy explored here. This move is likely to bypass the scalability problem and would allow to reconsider the paradigm of very large time steps without having to pay a high computational price for frequent spectral transforms.

However, the spectral-transform method did also provide a bunch of attractive features which must be kept in mind when building a grid-point alternative: (i) very high accuracy in the calculation of derivatives on a non-staggered horizontal grid (A-grid); (ii) trivial swap from differential vector fields (divergence, vorticity) to physical vector components (zonal and meridional velocities); (iii) trivial access to implicit numerical diffusion; and finally (iv) direct and highly-accurate solver for the implicit problem.

Some aspects related to these features in view of non-spectral models are now examined. Implications of using A-grids in numerical models based on grid-point algorithm have been extensively discussed in the literature where it has been noted that this choice leads to a spurious stationarity for shortest resolved gravity waves (see for example Mesinger (1979)). However, as mentioned in Bénard and Glinton (2019), this statement only applies to space-discretizations with low-order accuracy. For higher accuracy orders (e.g. 4 to 8) the problem is in practice relegated to the end of the resolved spectrum which, in any case, must be filtered out to avoid quadratic aliasing. Some current grid-point-based models appear to use an unstaggered A-grid without problems (e.g. Kühnlein et al. (2019)). Point (ii) may be disregarded in the first instance, when dealing only with idealised flows in the vertical plane: in this case, no swap between wind divergence and wind velocity is used (the model may be formulated in terms of wind components). In such an idealised model, the implementation of an implicit numerical diffusion is also not a critical issue; moreover, most of the cases presented here pertain to linear regimes and do not require diffusion. Finally, the main advantage lost here with a grid-point SI model, is the existence of an efficient and accurate direct solver. The main focus of the study is therefore the convergence of the iterative solver, since this drives almost entirely the overall performance of the SI-SL grid-point method.

Because of the large dimension of the implicit space problem, typically involving 10^9 variables, but thanks to its sparsity, several alternatives to direct solvers are extensively discussed in the literature (e.g. Müller and Scheichl (2014)). For instance, multigrid methods have been explored and

may be slightly faster than FFT-based algorithm as shown in Hess and Joppich (1997) with an Helmholtz equation in spherical coordinates from a local NWP model. Krylov methods are also used for example in Smolarkiewicz and Margolin (1994) where the Generalized Conjugate Residual method (GCR) is used successfully in a density-stratified potential flow past a steep three-dimensional isolated hill on a plan or in Thomas et al. (1997) where a preconditionned Generalized Minimal Residual Method (GMRES) solver is successfully used in a global SI-SL model, very close to AROME model. All these Krylov iterative methods seem particularly well suited since they mainly require local operations (as in HEVI schemes), the number of which depends on the accuracy required and the speed of convergence to achieve it. Some global calculations are still needed because of scalar products involved in the method, but these products need not be performed extensively if convergence is fast enough. Some studies as e.g. Zheng and Marguinaud (2018) suggest that the communication cost of scalar products can be more penalizing than the local communications if many iterations are required in the solver and more than 10^5 computing nodes are used. However, current models are typically operated on 10^2 nodes. It can therefore be assumed that neglecting the cost of scalar products will be relevant for the next few years, though it might become an issue in some future. Since the cost problem of scalar products is considerably alleviated if convergence is achieved with few iterations, the study of the actual speed of convergence in realistic contexts –as examined here– is a crucial aspect.

When only the most general characteristics of the problem are known (symmetry, definite positiveness etc), convergence bounds are then too loose to provide a usable estimate of the speed of convergence, but the knowledge of these characteristics may favour a particular Krylov method among the wide variety of those available (GMRES, GCR, etc). For instance, Steppeler et al. (2003) highlights the importance of the appropriate choice of the method for a non-symmetric linear problem.

Then, it becomes possible to conduct informative scalability tests or to test different preconditioners outside the constrained framework of an actual model (Müller and Scheichl 2014). However, these results are not necessarily generalisable because most of the problems inverted in meteorology depend strongly on the model itself: Poisson problem for most filtered systems (e.g. for the anelastic approximation), Helmholtz problem for implicit discrete systems, with variants depending on the assumptions made to construct the linear operator (presence of the orography or not,...), or the type of grid used, etc. Thus, convergence estimates obtained from a given framework are not necessarily transposable to another one, and require specific studies like in Skamarock et al. (1997) or in Qaddouri and Lee (2010). However, no operational NH model except AROME, simultaneously combines a SI-SL dynamic using an A-grid with a mass coordinate, thus making necessary a specific study.

Once the stopping criterion has been defined, the critical threshold at which the solver loop is exited remains to be defined. This tuning is essential because an advantage of using a Krylov method is that it converges towards an acceptable solution (in a sense defined below) in only a few iterations in most cases. This threshold is usually defined in an academic framework, like this one, and then generically applied in an operational model. Two approaches can be considered to define it at a given time of a simulation, by measuring the deviation of the solution of the partially-converged solution computed by the solver from:

- the converged solution of the problem (with an error accuracy close to the finite precision of floating-point arithmetic);
- a reference solution.

On the last point, except for a few linear flows, analytical solutions are generally not known. Consequently, the reference solution chosen in this study is the one provided by the spectral version of the AROME model, a model that has now been validated for several years and will be regarded in this study as a reference in terms of quality. Since most of the studies are carried out with only one model, this second approach is particularly original and consolidates the approach.

As mentioned in Liesen and Tichý (2004) the convergence rate does not depend only on the eigenvalue spectrum of the problem, but also on : the initialization of the iterative process; and the spectral properties of the right-hand side (RHS): in our case the meteorological situation. All these points must be carefully studied to obtain the best convergence properties at a given quality.

The last two points strongly suggest to test convergence in a real model: this makes possible to initialize the solver from the solution at the previous time step, instead of starting from arbitrarily-prescribed meteorological fields.

These points lead to build a full grid-point version of AROME using a Krylov solver. The aim of this study is to assess convergence properties of this solver in an idealized but realistic meteorological framework. Since no qualitative differences in convergence are expected to occur between 3D and two-dimensional (2D) vertical-plane simulations, or between idealized and real-case experiments, results are assessed through idealized 2D test cases, drawn from the literature. In practice the 2D grid-point version of the model is derived from the pre-existing 2D (vertical plane) spectral version of AROME. This spectral version, giving the same response as would the operational model, is used henceforth to produce the aforementioned reference solutions.

Because improvement of models in the future will not necessary involve refining the resolution at the same rate as what has been done so far (see e.g. Wedi (2014)), assessments in this study will be carried out in two complementary situations: one where the time step is adjusted at each change of resolution to keep the ratio time step over horizontal resolution constant, and one where the time step decreases at a given horizontal resolution, to better represent orographic effects. All these experiments will enable scalability estimates to be deduced in comparison to a HEVI model, regarded as a reference in terms of scalability.

In section 2, the equations and temporal discretization of the AROME model are written in a synthetic two-dimensional (vertical plane) formalism, reflecting the models used here to perform some idealized test case experiments extensively studied in the literature. Namely, the current spectral dynamic kernel (SP hereafter) is

described there and the study of its properties allows to build its grid-point (GP hereafter) counterpart benefiting from the best eigenvalue properties of the implicit problem to be inverted. Test cases and numerical configurations used are presented in section 3. In section 4, a quality evaluation is lead to validate the grid-point converged and spectral solutions. Then, the critical threshold at which the solver loop is exited is tuned so that the errors made compared to the spectral version are sufficiently small in comparison to those committed in an operational context. In section 5, a synthetic metric is defined to derive scalability estimates relative to an HEVI model in the two complementary situations previously mentioned. Finally, a conclusion and some perspectives are presented in section 6.

2. Model formulation

2.1. Governing equations

As already mentioned, the 2D vertical plane model used in this study is derived from the operational 3D model AROME (Bénard et al. 2010): fully compressible Euler Equations are formulated with a hybrid mass-based terrain-following coordinate η (Laprise 1992), and the transport scheme is based on the semi-Lagrangian technique.

The time evolution of the state-vector X gathering all prognostic variables may be written symbolically as:

$$\frac{dX}{dt} = \mathcal{M}(X) + \mathcal{F}(X), \tag{1}$$

where $\mathcal{F}$ is the physical sources term, $\mathcal{M}$ is the complete dynamical core model (see Appendix A and Appendix B). In this formalism, X is therefore a vector of differentiable functions in space and time. For further convenience, the state-vector X is split into two parts $X = [U, \Psi]^T$, with $\Psi = [d, T, \hat{q}, \pi_s]^T$ and where U is the horizontal wind velocity, d the reduced vertical divergence, T the temperature, $\hat{q}$ the reduced non-hydrostatic pressure, and π_s the hydrostatic surface pressure. Unlike in the 3D model, the zonal wind U is used here instead of horizontal divergence of the wind D. Consequently some second-order horizontal derivatives in the original 3D version of $\mathcal{M}$ are replaced here by first-order derivatives.

The linear system $\mathcal{L}$ used for the implicit scheme is derived from the above complete system, through some linearization around a stationary reference state X^*, with no orography, resting, hydrostatically-balanced, isothermal and dry. The resulting linear system involves a reference hydrostatic surface-pressure π_s^* and two reference temperatures T^* and T_e^* (Bénard, 2004). The expression of $\mathcal{L}$ is similar to its 3D counterpart (cf Bénard, et al. 2010):

$$\begin{aligned}
\partial_t U &= -R_d\mathcal{G}^*\partial_x T + R_d T^*\mathcal{G}^*\partial_x\hat{q} - R_d T^*\partial_x\hat{q} \\
&\quad - \frac{R_d T^*}{\pi_s^*}\partial_x\pi_s, \\
\partial_t d &= -\frac{g^2}{R_d T_e^*}\mathcal{L}^*\hat{q}, \\
\partial_t T &= -\frac{R_d T^*}{C_{vd}}\left(\partial_x U + d\right), \\
\partial_t\hat{q} &= \mathcal{S}^*\partial_x U - \frac{C_{pd}}{C_{vd}}(\partial_x U + d), \\
\partial_t\pi_s &= -\pi_s^*\mathcal{N}^*\partial_x U,
\end{aligned} \tag{2}$$

where $\mathcal{G}^*$, $\mathcal{L}^*$, $\mathcal{S}^*$ and $\mathcal{N}^*$ are vertical operators defined in Appendix C. Since all the coefficients in (2) are prescribed and kept constant in time and space, the implicit scheme is a 'constant-coefficient' one, in the sense discussed above.

2.2. Time discretization

As detailed in Bénard et al. (2010), Euler equations are integrated by an Iterative Centered Implicit (ICI) scheme

$$\begin{aligned}
\frac{X^{+(n)} - X^0}{\Delta t} &= (\mathcal{M} - \mathcal{L})\left(\frac{X^{+(n-1)} + X^0}{2}\right) \\
&\quad + \mathcal{L}\left(\frac{X^{+(n)} + X^0}{2}\right) + \mathcal{F}(X^0),
\end{aligned} \tag{3}$$

where Δt is the time step and the superscript $+(n)$ denotes the future time level obtained after the n-th iteration of the ICI scheme. Currently, in the operational context only 2 iterations are required. This configuration is sometimes called a 'predictor-corrector' scheme. At each iteration of the scheme, the implicit problem to be solved may be written as

$$\left[\mathcal{I} - \frac{\Delta t}{2}\mathcal{L}\right] X^{+(n)} = X^\bullet, \tag{4}$$

where:

$$\begin{aligned}
X^\bullet &= X^0 + \Delta t(\mathcal{M} - \mathcal{L})\left(\frac{X^{+(n-1)} + X^0}{2}\right) \\
&\quad + \Delta t\mathcal{L}\left(\frac{X^0}{2}\right) + \Delta t\mathcal{F}(X^0),
\end{aligned} \tag{5}$$

or equivalently, using a block matrix formalism for (2)

$$\left[\begin{array}{c|c} \mathcal{I} & -\frac{\Delta t}{2}\mathcal{L}_A \\ \hline -\frac{\Delta t}{2}\mathcal{L}_B & \mathcal{I} - \frac{\Delta t}{2}\mathcal{L}_C \end{array}\right] \left[\begin{array}{c} U^{+(n)} \\ \hline \Psi^{+(n)} \end{array}\right] = \left[\begin{array}{c} U^\bullet \\ \hline \Psi^\bullet \end{array}\right]. \tag{6}$$

In (6), $\mathcal{I}$ denotes the identity operator in the relevant space, $\mathcal{L}_A = \mathcal{V}_A \circ \partial_x$, $\mathcal{L}_B = \mathcal{V}_B \circ \partial_x$, and

$$\mathcal{L}_C = \left[\begin{array}{c|c|c|c} 0 & 0 & -\frac{g^2}{R_d T_e^*}\mathcal{L}^* & 0 \\ \hline -\frac{R_d T^*}{C_{vd}} & 0 & 0 & 0 \\ \hline -\frac{C_{pd}}{C_{vd}} & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 \end{array}\right], \tag{7}$$

with

$$\mathcal{V}_A = \left[\begin{array}{c|c|c|c} 0 & -R_d\mathcal{G}^* & R_dT^*\mathcal{G}^* - R_dT^* & -\frac{R_dT^*}{\pi_s^*} \end{array}\right], \tag{8}$$

$$\mathcal{V}_B = \left[\begin{array}{c} 0 \\ \hline -\frac{R_dT^*}{C_{vd}} \\ \hline \mathcal{S}^* - \frac{C_{pd}}{C_{vd}} \\ \hline -\pi_s^*\mathcal{N}^* \end{array}\right] \tag{9}$$

Operators $\mathcal{L}_A$ and $\mathcal{L}_B$ contain both vertical ($\mathcal{V}_A$ and $\mathcal{V}_B$) and horizontal (∂_x) operators and $\mathcal{L}_C$ is a contribution of the non-hydrostatic vertical part, reduced to zero in the hydrostatic version. The system (6) may then be reduced to a single variable equation by an appropriate algebraic elimination in favour of U

$$\left[\mathcal{I} - \frac{\Delta t^2}{4}\mathcal{B}\circ\partial_x^2\right] U^{+(n)} = U^{\bullet\bullet} \tag{10}$$

where $\mathcal{B}$, referred to as the non-symmetric vertical operator, is defined by

$$\mathcal{B} = \mathcal{V}_A\left(1 - \frac{\Delta t}{2}\mathcal{L}_C\right)^{-1}\mathcal{V}_B,$$

and

$$U^{\bullet\bullet} = U^{\bullet} - \frac{\Delta t}{2}\mathcal{L}_A\left(1 - \frac{\Delta t}{2}\mathcal{L}_C\right)^{-1}\Psi^{\bullet}.$$

2.3. *Space discretization*

The time-discrete model described hitherto involved space-continuous variables and operators. A space-discretized version is then built, which detailed presentation is out of the scope of this paper. The model used here being indeed a 2D clone of the operational 3D version of AROME, the discretization follows Bénard et al. (2010) except for the minor changes implied by the use of U instead of D. The only important feature to be highlighted here is the vertical discretization ensures that the implicit problem created by (3) may still be reduced to a single variable problem as in (10). This requires some appropriate properties of vertical operators, always valid in the continuous context, to be also satisfied for vertically-discrete operators. In the vertical discretization of AROME, this constraint is satisfied (see Bubnová et al. (1995)), thereby permitting the above algebraic elimination procedure to hold in the space-discrete case.

For notation convenience, all symbols used so far to describe space-continuous variables and operators are kept the same to describe space-discretized variables and operators henceforth. An equation formally similar to (10) is therefore arrived at, but now involving space-discrete variables and operators. The aim of this paper is therefore to examine and compare various ways to solve this pivotal equation. Regarding this equation, the difference between the spectral and grid-point versions of the 2D model only lays in the concrete form of the discrete operator ∂_x^2.

2.4. *Decomposition into the eigenspace of $\mathcal{B}$*

The total number of levels in the vertical discretization is denoted N_{lev}. In a 2D framework, the space-discrete version of (10) implies a large-size 2D implicit problem. A possible alternative approach is to exploit the separability of this equation along horizontal and vertical directions, and to decompose this problem into a set of N_{lev} two-dimensional problems by projecting the space-discrete version of (10) in the eigenspace of the vertical operator $\mathcal{B}$. This directional separation in the implicit problem is a *raison d'être* of the spectral technique and is therefore always used in the spectral version of the model. For grid-point versions, both approaches could equally be considered, but the potential advantage of exploiting the separability is indeed used in this paper (see related discussion in section 2.6).

For a given eigenmode $\mathcal{B}_l$ of the operator $\mathcal{B}$, the 1D implicit equation to be solved is

$$\left[1 - \frac{\Delta t^2}{4}b_l\,\partial_x^2\right]\widetilde{U}_l^{+(n)} = \widetilde{U}_l^{\bullet\bullet}, \tag{11}$$

where b_l is the l-th eigenvalue of $\mathcal{B}$ sorted in descending order, and ($\widetilde{U}_l^{+}$, $\widetilde{U}_l^{\bullet\bullet}$) are the projected component of (U^{+} $U^{\bullet\bullet}$), on the corresponding eigenmode $\mathcal{B}_l$.

$\mathcal{B}$ can be approximately seen as the product of an inverse operator on the vertical by the square of the sound speed. For the most external mode b_0, sometimes called the 'barotropic' mode, the propagation is essentially horizontal (and the part corresponding to the inverse operator tends towards identity) while for the internal modes, or 'baroclinic' modes, the propagation is essentially vertical (the part corresponding to the inverse operator tends towards zero).

The spectrum of $\mathcal{B}$ is plotted in Fig. 1 for various time steps Δt in the non-hydrostatic (NH) and hydrostatic (H) cases, for an horizontal grid-mesh of $100\,\text{m}$ and the vertical grid used in all experiments below. Eigenvalues are positive regardless of Δt, and rapidly decrease toward zero with the vertical mode index. This property will be strongly exploited in the following. It can be noted that the eigenvalues of $\mathcal{B}$ are strongly dependent on the time step for the most internal modes but not for the most external ones where vertical propagation is weak. The representation of vertical wave propagation is improved as the time step decreases. On the contrary the eigen-spectrum of $\mathcal{B}$ of the NH model tends towards which of the H one, as the time step increases. Thus, the improvements of the vertical wave propagation (especially gravity waves) sought during the implementation of an NH model in favour of an H model, are fully satisfied when the time step is small, as previously mentioned.

For a given resolution $(\Delta x, \Delta t)$, an horizontal wave-CFL number associated to the l-th mode may be introduced as

$$c_l^* = b_l^{1/2} \frac{\Delta t}{\Delta x}. \quad (12)$$

The maximum eigenvalue b_0 is the square of the sound speed and the corresponding value c_0^* is the CFL number for horizontally-propagating sound waves. Since this is the most constraining value in the whole set of horizontal CFL numbers c_l^*, c_0^* will be used as the reference wave-CFL number in each of the following experiments.

2.5. *Spectral version*

In the 2D model, a Fourier representation of space-discretized variables along x is possible by making all fields periodic through an appropriate extension technique (Haugen and Machenhauer (1993)). The extended vectors U and $U^{\bullet\bullet}$ may then be projected into the Fourier space using an efficient FFT algorithm, the decomposition $\widehat{U}$ of a field U being termed the 'spectral transform' of U. Since Fourier components $\exp(ikx)$ are the eigenmodes of the horizontal periodic Laplacian operator ∂_x^2, the implicit problem (10) projected in Fourier space is reduced to scalar inversions

$$\left[1 + \frac{1}{4} c_l^{*2} k'^2\right] \widehat{\widetilde{U}^{+(n)}} = \widehat{\widetilde{U}^{\bullet\bullet}} \quad (13)$$

where $k' = k\Delta x \in [-\pi, \pi]$ is the non-dimensional zonal wavenumber.

2.6. *Grid-point version*

Choice of strategy

Once the horizontal operator ∂_x^2 is discretized on the A-grid, one of the following strategies has to be chosen:

i. either solving N_{lev} 1D independent symmetric problems (11);
ii. or solving the large 2D non-symmetric problem (10).

As mentioned previously, the convergence of a Krylov solver depends on the eigenvalues spread of the linear operator matrix, measured by its condition number: the higher it is, the slower the convergence is. When a sufficiently high order is used to discretize the one-dimensional Laplacian operator so that its response is close to the spectral one and considering the vertical operator $\mathcal{B}$ is almost symmetric, the condition number is then reduced to:

i. $C_l \simeq 1 + \frac{\pi^2}{4} c_l^{*2}$, for the N_{lev} independant 1D symmetric problems (11);
ii. or $C \simeq 1 + \frac{\pi^2}{4} c_0^{*2}$, for the large 2D non-symmetric problem (10) (that is to say the full 2D system is as ill-conditioned as the worst 1D equation).

The fastest vertical mode, although being the most difficult to solve, has a relatively low condition number, $C_0 \simeq 500$ (corresponding to the current operational configuration), when compared to typical values encountered in recent literature (Ye (2017), Soleymani (2013)) with typical values of ill-conditioned matrices larger than 10^{10}. The remaining $N_{i,i=1\ldots N_{\text{lev}}}$ problems are even better-conditioned and a very fast convergence using any iterative grid-point solver is expected for the majority of them.

To exploit as best as possible the current paradigm of parallelization, one has a strong interest to treat the N_{lev} vertical modes sequentially instead of treating them in parallel and waiting for the barotropic mode to finish, as it would be the case by solving the large non-symmetric problem (10). Consequently solving the N_{lev} independant symmetric problems will be the chosen method. Because of the positivity of the b_l modes, all these problems are symmetric and positive definite which allows using specific algorithms like the conjugate gradient one. For the following, it is more convenient to write the lth problem under the form :

$$\mathcal{H}x = y \quad (14)$$

where : $\mathcal{H} = \left[1 - \frac{1}{4} c_l^{*2} \delta_x^2\right]$, $x = \widetilde{U}_l^{+(n)}$ and $y = \widetilde{U}_l^{\bullet\bullet}$.

Initialization

An interesting approach to improve the rate of convergence is to initialize the solver by an estimate of the atmospheric state x_0 as close as possible to the sought solution y. The best way we found is by taking the field at the previous time step or at the previous iteration of the ICI scheme (not shown). The solution (of each of the N_{lev} problems) is then looked for in the subspaces of (different) size N :

$$\mathcal{K}_N = \left\{r_0, \mathcal{H}r_0, \mathcal{H}^2 r_0, ..., \mathcal{H}^N r_0\right\} \quad (15)$$

where $r_0 = y - \mathcal{H}x_0$ is the initial residual.

Stopping criterion

The linear system being symmetric, thus, according to Proposition 4 in Saad and Schultz (1986) the following inequality holds:

$$\frac{||r_n||}{||r_0||} \leq \left(1 - \frac{1}{C_l^2}\right)^{n/2} \quad (16)$$

where the residual at iteration n is $r_n = y - \mathcal{H}x_n$. The most obvious and widely used stopping criterion with such methods is to halt the iterative process when the residual value is reduced by a certain amount compared to the norm of the RHS:

$$\text{Stop if } \frac{||r_n||}{||y||} \leq \varepsilon \quad (17)$$

where $\varepsilon < 1$ is an a-priori reduction threshold fixed by the user. That inequality is the worst case convergence inequality and in practise the convergence rate can be much faster (Axelsson (1996)). In this study we assume ε does not depend on the vertical mode considered. From equation (16) it can be deduced that the number of iterations of the algorithm n varies as the logarithm of the residual parameter ε in (17). Consequently it is relevant to choose what we call 'the residual reduction parameter' ε^{-1} with values of the form 10^p in the numerical experiments to follow.

3. Presentation of test cases

The following benchmark composed of two cases has been widely used for model validation purposes and thus results can be compared to existing simulations. It has been designed to be a challenge for meso-scale forecasting which intends to model convection and gravity waves accurately involving advection and adjustment terms respectively.

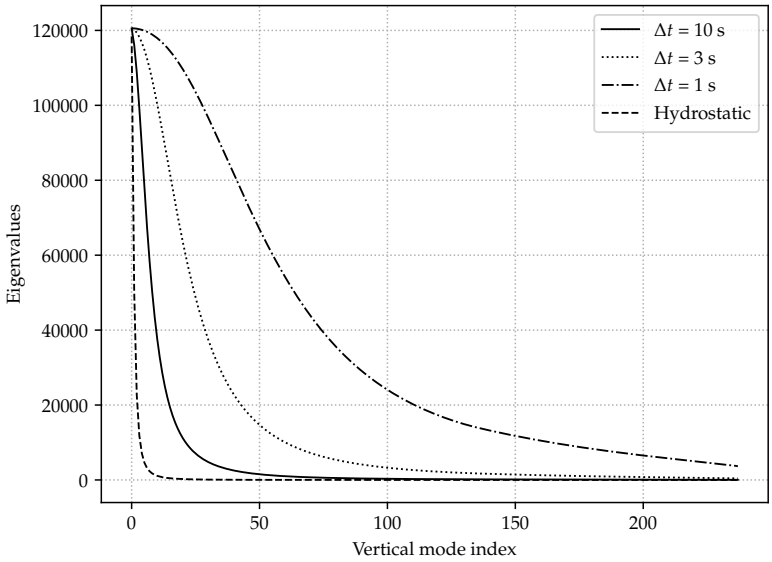

Figure 1. Eigen-spectrum of $\mathcal{B}$ with different time step under an horizontal resolution of 100 m and the vertical grid configuration described section 3.3. The spectrum of the hydrostatic version which is independent of the time step is also plotted.

3.1. Warm bubble test case

The warm bubble test case simulates the evolution of a buoyant thermal in a constant potential temperature environment of 300 K. The version used here is described in Carpenter Jr et al. (1990) where a background uniform speed of 20 $\mathrm{m{\cdot}s^{-1}}$ like in Wicker and Skamarock (1998) is added. A warm anomaly of radius 2 km and centred at 2 km height placed at the middle of the horizontal is given at the beginning. During the simulation the warm bubble rises and develops while being advected. The simulation ends after 1000 s.

3.2. Orographic test case

The orographic test case used in Schär et al. (2002) simulates gravity waves in a stratified environment over a mountain profile given by:

$$h(x) = h_{\max} e^{-(x/r_{\mathrm{a}})^2} \cos^2(\pi x / r_{\mathrm{b}}) \qquad (18)$$

where $h(x)$ is the orography profile, $h_{\max}$ = 250 m is the maximum orography height, r_{a} = 5000 m is the overall mountain range width and r_{b} = 4000 m is the length scale of the sub-orographic modulation. The initial condition is an homogeneous wind profile of 10 $\mathrm{m{\cdot}s^{-1}}$ with a ground temperature of 288 K in a constant Brunt-Väisälä frequency of 0.01 $\mathrm{s^{-1}}$ environment. The simulation ends after 8000 s, once a near stationary state is reached.

3.3. Configuration

In both cases an horizontal domain size of 100 km and a vertical height of 10 km is given. Except when explicitly mentioned, the horizontal resolution Δx is 100 m and vertical levels are chosen such that for a reference vertical temperature profile and surface pressure, the vertical resolution is 100 m up to 10 km. Above, the vertical levels have a constant pressure depth up to the top of the domain that corresponds to a 0 hPa surface.

In the following experiments and for time steps larger or equal to 3 s, an horizontal numerical diffusion is applied at the end of the time step to overcome the ill-represented energy dissipation at the smallest scales. It is performed by a filtering applied to each wavenumber in spectral whereas a fourth order diffusion (such as ∇^4) is applied in grid-point.

In the following, a fourth-order finite difference derivative scheme has been used. Indeed, sensitivity experiments have proven that it is noticeably better than a second order scheme, whereas a sixth order scheme showed very little improvement. In addition, with a view to possibly solving a more general (non-symmetric) problem in the future, e.g. by including orographic terms in the implicit problem, experiments are conducted using the GMRES algorithm.

4. Quality evaluation

In this section, the converged solution of the grid-point model is compared to the spectral one and to the literature. It is shown they are very close to each other and thus allowing to define both of them as references. Then, the critical threshold ε is adjusted and the resulting solutions are compared to these references and lead to define a relevant threshold for the stopping criterion.

4.1. Comparison with the spectral version and literature

The spectral (SP) and grid-point (GP) model forecasts for the rising bubble case are shown here. The converged GP solution is obtained by setting a non optimal very large value for ε^{-1} leading to an error accuracy close to the finite

precision of floating-point arithmetic. Results are presented for various CFL number (c_0^*) values:

$c_0^* = 60$

Here Δt = 20 s, and with this very large CFL number it is interesting to notice that the scheme is nevertheless stable, but as expected because of the long time step, the shape of the solution (not shown) is of poor quality with a maximum temperature of 301.64 K in good agreement with the Δt = 2 s solution (see below the paragraph corresponding to $c_0^* = 6$) but with a much too low maximum vertical velocity only of 7.5 $\mathrm{m{\cdot}s^{-1}}$.

$c_0^* = 9$

The case with Δt = 3 s corresponding to a $c_0^* = 9$ value comparable to the one used in operational configuration has been reproduced. Figure 2 shows the comparison between SP and GP versions after 1000 s. As explained above, diffusion is treated differently in the two models, this is the largest source of difference in Figure 2. The shape of the bubble is also not perfectly symmetric to the median axis contrary to what would be expected. This is due to the background horizontal velocity field that makes the rising thermal more complex to simulate. Both GP and SP solutions are very close for the potential temperature and the vertical wind speed (not shown) with a maximum temperature of 301.83 K for the SP simulation and 301.77 K for the GP one and 13.03 $\mathrm{m{\cdot}s^{-1}}$ and 12.39 $\mathrm{m{\cdot}s^{-1}}$ respectively for the vertical wind.

$c_0^* = 6$

Figure 3 presents results of the same experiment but with a time step of 2 s. Now the solution is much more symmetric and similar to the numerical solution obtained with other models in literature using the same time step: the maximum vertical velocity in the rising thermal is 13.3 $\mathrm{m{\cdot}s^{-1}}$ to be compared with 16.5 $\mathrm{m{\cdot}s^{-1}}$ in Wicker and Skamarock (1998) and 14 $\mathrm{m{\cdot}s^{-1}}$ in Bryan and Fritsch (2002). The maximum potential temperature of 301.85 K, which is very similar to the values found in Wicker and Skamarock (1998) and Bryan and Fritsch (2002). Here when compared to Figure 2 since no diffusion is used, SP and GP simulations are closer to each other.

From the result presented in this section, it seems that the converged GP solution is very close to the SP one, whatever the CFL number c_0^* is. Both are also very close to what is obtained in the literature provided $c_0^* \leq 6$. This confirms the relevance of keeping the main characteristics of the model and especially the use of the A-grid with a sufficiently high-order space-accurate scheme. Moreover, the converged GP solution or the SP solution can both be defined as references in terms of quality.

4.2. *Stopping criterion*

Contrary to the previous section, the quality of the solution is now degraded by changing the values of the residual reduction ε^{-1} and deviation from references solutions are measured via the root mean square error (RMSE) on the potential temperature and vertical wind speed fields according to the test case used. These results are plotted in Figure 4.

Warm bubble test case

It can be shown first that as expected, the larger the residual reduction is, the closer the solution is from the exact grid-point solution (dotted line on Figure 4). However there is no plateau or inflection of the RMSE's curve that would be helpful to retrieve an optimal residual reduction ε^{-1} beyond which further iterations are useless.

The RMSE also decreases when the solution of GP is compared to SP for residual reduction up to 10^3. Then, beyond that value, the error due to the use of an approximated iterative solution is less than the difference between the GP and the SP model so that the extra iterations do not bring the GP solution closer to the SP one. The RMSE between the converged GP solution and the SP one cannot therefore go below 0.036 K for the potential temperature and 0.18 $\mathrm{m{\cdot}s^{-1}}$ for the vertical wind (RMSE's corresponding to $\varepsilon^{-1} = 10^3$).

Furthermore, the order of magnitude of RMSEs for the current spectral version of AROME in an operational forecast is about 1 $\mathrm{m{\cdot}s^{-1}}$ for wind and 1 K for temperature even for short time ranges less than one hour (see for example Auger et al. (2015)). Therefore these RMSEs are significantly lower to other sources of model error. Considering a RMSE for temperature less than 0.05 K (lower by a factor about 20 compared to errors made in an operational forecast) is still an acceptable error, a residual reduction of $\varepsilon^{-1} = 10^2$ is sufficient and will be taken as an acceptable threshold for the warm bubble test case.

Orographic test case

In a similar way, the maximum acceptable RMSE threshold for the different experiments was set to 0.01 $\mathrm{m{\cdot}s^{-1}}$ by trial and checking that the gravity waves patterns are reproduced as well as for the reference such as the ones to be seen on Figure 5.

To conclude, we managed to carefully define a threshold for the stopping criterion so that the errors committed by a solution that is not fully converged is significantly lower (at least 20 times) than the errors traditionally committed in an operational context. This estimate will be applied to the scalability study in the following section.

5. Scalability evaluation

In this section, we first define a synthetic scalability criterion and then conduct several experiments in which we evaluate the communication cost by changing: the critical threshold for the stopping criterion, the spatial resolution keeping the $\Delta t/\Delta x$ ratio constant and then reducing it. In any cases, the same vertical grid is kept.

5.1. *A measure of scalability*

As mentioned in the introduction, an appropriate, scalable alternative to the SI-SL approach is the HEVI time stepping method since it requires a minimum transfer of data (i.e.

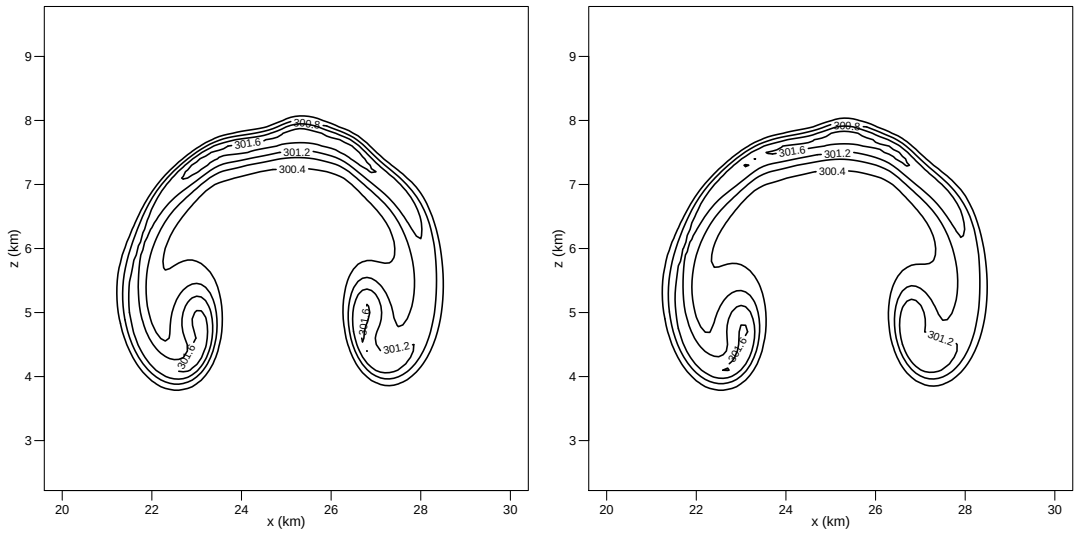

Figure 2. Rising thermal simulation with $\Delta x = 100\,\mathrm{m}$ and $\Delta t = 3\,\mathrm{s}$ after $1000\,\mathrm{s}$: (a) potential temperature for the GP version; (b) potential temperature for the SP version.

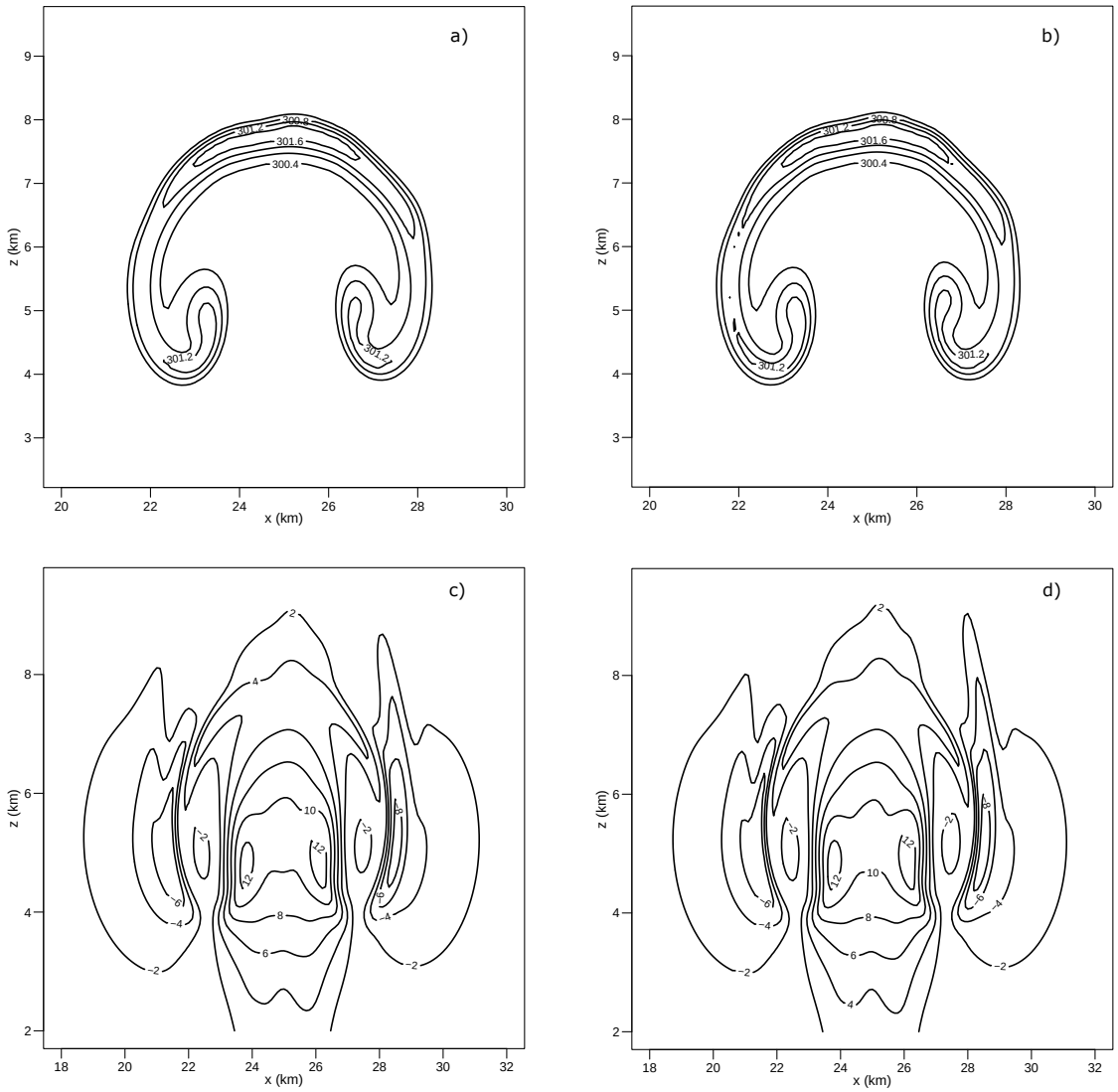

Figure 3. Rising thermal simulation with $\Delta x = 100\,\mathrm{m}$ and $\Delta t = 2\,\mathrm{s}$ after $1000\,\mathrm{s}$: (a) potential temperature for the GP version; (b) potential temperature for the SP version; (c) vertical wind speed for the GP version; (d) vertical wind speed for the SP version.

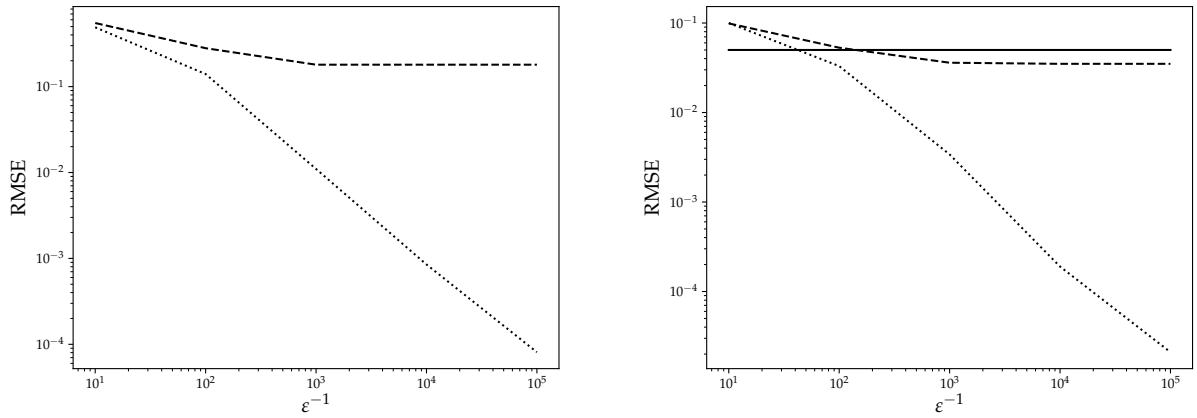

Figure 4. RMSE compared to the SP version (dashed line), the exact GP version (dotted line) for the warm bubble test case as a function of the ε^{-1} criterion with $\Delta x = 100\,\mathrm{m}$ and $\Delta t = 2\,\mathrm{s}$ after 1000 s. Left: vertical velocity field; right: potential temperature field. Acceptable threshold for a typical temperature forecast is shown in continuous line.

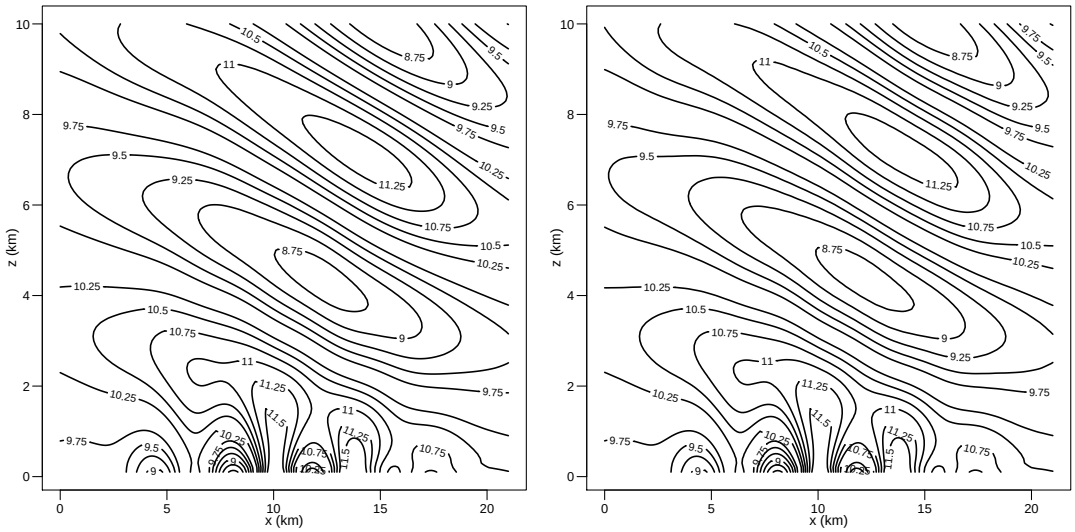

Figure 5. Mountain waves simulation with Δx = 100 m and Δt = 3 s, horizontal wind after 8000 s. Left: reference for the GP version; right: with a residual reduction ε^{-1} of 300.

only a few layers surrounding the sub-domain owned by a computer node). Therefore it is interesting to evaluate the communication efficiency by looking at the amount of data exchanged in comparison to a typical HEVI scheme which will be briefly described in the following.

Recent HEVI models often combine an implicit solver for rapid waves on the vertical, an accurate explicit treatment of slow terms (e.g. Runge-Kutta integration) and a cheap explicit treatment for the horizontal propagation of fast waves (Gassmann (2013), Durran (2013)). Such a split-explicit scheme uses a short time step also named acoustic time step ($\Delta\tau$) which, in order to ensure numerical stability, must satisfy the CFL constraint:

$$\frac{c_s \Delta\tau}{\Delta x} \leq \alpha \tag{19}$$

where c_s is the sound velocity, $\alpha \simeq 1/\sqrt{2}$ is a factor that depends on the order of the temporal and spatial schemes used, the grid used, etc. Now the assumption will be made that, in an operational implementation, most of the inter-node communication is due to the small time step integration for the HEVI scheme, and due to the iterative semi-implicit inversion for AROME. Thus, during the acoustic time step integration the prognostic variables are updated following:

$$X^{t+(n+1)\Delta\tau} = \mathcal{H}' X^{t+n\Delta\tau} \tag{20}$$

where $X^{t+n\Delta\tau}$ is the state vector at the small time step $t + n\Delta\tau$, t is the previous long time step and $\mathcal{H}'$ has a similar computational cost compared to $\mathcal{H}$ in (14), representing also the terms controlling the horizontal propagation of the fast waves. Then to make a forecast until $t + \Delta t$, where Δt has the same value as a SI time step, several applications of iteration (20) are required and after $N^{'}$ small time steps, the solution is:

$$X^{t+N^{'}\Delta\tau} = \mathcal{H}'^{N^{'}} X^{t} \tag{21}$$

There are similarities between HEVI and Krylov methods used in a SI forecast model. They both involve applying several time linear operators that are similar in terms of computation cost but also in terms of exchange of data. Since most of the communication cost comes from the data exchanges for horizontal derivatives, the application of $\mathcal{H}'$ will behave like $\mathcal{H}$ scalability-wise. Then the scalability problem can be reduced to the comparison between the values N and N' or equivalently by the effective time steps of both methods. We define the effective or equivalent explicit time step Δt_{eq} of a SI Krylov method as the time step divided by the overall number of iterations vertically averaged because of the eigenmode decomposition. An explicit model with this time step would perform similarly in terms of scalability. For example AROME which uses a predictor-corrector scheme to enable the use of longer time step, necessitates two stages at each time step leading to solving twice the implicit scheme using the solver. Consequently:

$$\Delta t_{eq} = \frac{\Delta t}{2\bar{N}} \tag{22}$$

where $\bar{N}$ is the average number of iterations required through all vertical levels. Then this equivalent time step can be compared to the accoustic time step of an HEVI model which can be guessed using the CFL constraint : for example in the context of the test cases presented here where $\Delta x = 100\,\mathrm{m}$, considering a sound velocity c_s of $350\,\mathrm{m{\cdot}s^{-1}}$, the reference small time step is then equal to $\Delta\tau \simeq 0.2\,\mathrm{s}$. That number agrees with published results, for instance, in Wicker and Skamarock (1998) a warm bubble test case was used with Δx = 125 m. Results shown were obtained with a second-order Runge-Kutta time integrator using a large time step of Δt = 2 s and a number of steps N' equals to 12 leading to $\Delta\tau \simeq 0.17\,\mathrm{s}$.

The scalability comparison factor can be introduced:

$$f = \frac{\Delta t_{eq}}{\Delta\tau} \tag{23}$$

as the ratio between the equivalent explicit time step and the HEVI acoustic time step. When $f > 1$, the SI method is more scalable than the HEVI one and reciprocally.

5.2. *Stopping criterion evaluation*

We intend in this part to assess how explicit equivalent time step and f evolve as a function of the residual reduction with the warm bubble test case.

As mentioned in Table 1 when the residual reduction ε^{-1} is greater than 10^3, the number of iterations of the solver increases such that the HEVI method ends up performing better than the SI-GP method ($f < 1$). When $\varepsilon^{-1} \simeq 10^2$, as advocated previously, the number of iterations obtained is such that the SI-GP method is a bit more scalable than the HEVI method ($f \simeq 1.45$) for a quality approximately equivalent according to observations made in section 4.1. This result also confirms the stopping criterion adjustement influences scalability performance.

5.3. *Future resolution keeping the ratio $\Delta t/\Delta x$ constant*

In an operational context up to now at least, the horizontal resolution is improved trying to preserve the ratio $\Delta t/\Delta x$ as much as possible for stability and accuracy reasons. Table 2 shows how the maximal and averaged number of iteration varies with a constant $\Delta t/\Delta x$ ratio similar to the one used with the operational version. Each experiment starts with a different bubble size adapted to its resolution and consequently forecast outcome is different from one experiment to another. The maximal number of iterations is obtained by respecting the RMSE threshold previously mentioned.

ε^{-1}	N_0	$\bar{N}$	Δt_{eq}	f
10^1	11.4	5.4	0.37	1.85
10^2	19.3	7.0	0.29	1.45
10^3	29.4	10.2	0.20	1.0
10^4	40.3	13.7	0.15	0.75
10^5	50.8	17.3	0.12	0.6

Table 1. Results of various experiments for the rising bubble experiment with $\Delta t = 2\,\mathrm{s}$ and $\Delta x = 100\,\mathrm{m}$. The maximum and averaged number of iterations per time step (N_0 and $\bar{N}$) with the associated equivalent explicit time step Δt_{eq} and comparison factor f for different residual reduction parameters ε^{-1} are shown.

Δt	Δx	ε^{-1}	N_0	$\bar{N}$	Δt_{eq}	f
30	1000	29	26.2	4.7	6.4	1.60
15	500	19	17.3	4.4	3.4	1.70
6	200	90	26.6	6.3	0.95	1.19
3	100	90	26.2	7.9	0.38	0.95

Table 2. Impact of the time step value keeping $\Delta t/\Delta x$ ratio constant in the warm bubble test case. The maximum and averaged number of iterations per time step (N_0 and $\bar{N}$) with the associated equivalent explicit time step Δt_{eq} and comparison factor f corresponding to the associated residual reduction under the RMSE threshold are shown.

Table 2 shows that the average number of iterations is roughly of the same order of magnitude and the comparison factor f is close to the unity whatever the values $(\Delta x, \Delta t)$ considered, thus confirming the good scalability. This also confirms the viability of keeping a constant $\Delta t/\Delta x$ ratio when improving spatial resolution.

However, the slight increase in the average number of iterations $\bar{N}$ (or equivalently the slight decrease of the comparison factor f) as $(\Delta x, \Delta t)$ decreases may come from the use of slightly different initial conditions on the one hand, and a slight degradation of the condition number of each of the N_{lev} problems due to the change in eigenvalue distribution of $\mathcal{B}$ when the time step decreases while the same vertical grid is kept (see for example Figure 1: eigenvalues are higher as the time step increases) on the other hand.

5.4. *Impact of time step reduction*

We intend in this part to assess how the explicit equivalent time step evolves as a function of the time step at constant horizontal resolution, with the same experimental configuration as previously mentioned (Δx =100 m) with the warm bubble and orographic test cases. For each time step, the residual reduction value shown was adjusted by trial and error to obtain a solution close enough to

the reference simulation just under the RMSE threshold imposed : see for example Figure 5 and Figure 6.

As mentioned in Table 3 and Table 4, the average number of iterations $\bar{N}$ decreases as the time step decreases. In the orographic case, a substantially higher residual reduction is required but results are very similar to what is obtained with the warm bubble test case.

In all cases tested the comparison factor f is greater than unity, meaning that the scalability of the method is maintained whatever the time step is. The slight decrease of the comparison factor f as the time step decreases comes from poorer condition numbers for each of the N_{lev} vertical problems, already mentioned in the previous paragraph. This is not a noticeable barrier since the comparison factor f remains larger than one even with smaller time step. Therefore, no obstacles have been identified from the scalability point of view if a reduction in the time step were to be implemented.

Δt	ε^{-1}	N_0	$\bar{N}$	Δt_{eq}	f
10	13500	87.2	18.2	0.55	2.75
5	3500	40.2	9.7	0.51	2.55
3	1100	27.7	7.6	0.39	1.95
2	300	19.1	6.6	0.30	1.50
1	100	9.4	5.0	0.2	1

Table 3. Impact of the time step value for the orographic wave test case.The maximum and averaged number of iterations per time step (N_0 and $\bar{N}$) with the associated equivalent explicit time step Δt_{eq} and comparison factor f corresponding to that residual reduction under the RMSE threshold are shown.

Furthermore, we notice that in the GP version, a decrease by a factor of 3 of the time step (going from 3 s to 1 s for example) does not lead to an increase by a factor of 3 of the communication cost as it would be the case in a spectral model (because to make a forecast at a given time, 3 times more spectral transforms steps would then be necessary, whereas the GP model benefits from a decrease of iteration per time step). An increase only by a factor barely more than 2 is measured: the equivalent explicit time step goes from 0.39 s to 0.2 s and from 0.38 s to 0.24 s respectively in the orographic (Table 3) and warm bubble (Table 4) test cases. Therefore, the overhead that occurs when the time step decreases at a given resolution turns out to be lower in the GP version than in the SP one. This comment, in addition to those already mentioned, enhances the attractiveness of the grid-point method.

Δt	ε^{-1}	N_0	$\bar{N}$	Δt_{eq}	f
20	270	106.7	14.9	1.34	6.7
10	190	70.4	12.7	0.79	3.95
5	150	43.3	10.3	0.48	2.40
3	90	26.2	7.9	0.38	1.90
2	55	17.6	6.1	0.33	1.65
1	4	5.8	4.2	0.24	1.20

Table 4. Impact of the time step value for the warm bubble test case. The maximum and averaged number of iterations per time step (N_0 and $\bar{N}$) with the associated equivalent explicit time step Δt_{eq} and comparison factor f corresponding to the associated residual reduction under the RMSE threshold are shown.

6. Conclusion

The objective of this paper was to present a scalable alternative to the current spectral AROME model while keeping most of its main characteristics (constant coefficient SI scheme, SL transport scheme, A-grid, mass based coordinate, etc). This alternative relies on the removal of spectral transforms to perform all calculations in grid-point space including the implicit problem, solved here by a Krylov algorithm. This class of algorithms consists in iterating local operators with few communications, the number of iterations depending on the speed of convergence and the desired accuracy.

The known convergence results are often too general to lead to a sufficiently precise estimate of the speed of convergence, so that a specific study is required with all the characteristics of the model under study. This work has been lead with a set of bidimensional non-linear flows which, though idealized, are reminiscent of realistic situations, and involve all source terms of the complete model.

The grid-point version developed here retains some beneficial features already present in the spectral version of the AROME model, namely the decomposition of the problem on the vertical eigenmodes, easily accessible through the choice of a 'constant coefficient' basic state.

The spectral version of the AROME model, which has been operational for several years, has been used here as a reference in terms of quality. In a first time, the possibility of producing forecasts of similar overall quality to spectral ones has been examined. It has found that for the converged solution (i.e. with ε close to the machine precision), grid-point forecasts are very close to spectral ones with the same grid, provided a sufficiently high-order space-accurate scheme is used (fourth-order and sixth-order gave very similar results whereas second-order resulted in poorer solutions). This result thus validates the strategy of keeping as much as possible the main characteristics of the current AROME model.

After correctly initializing the solver, the critical threshold at which the solver loop is exited has been adjusted to ensure that the errors between the partially-converged solution (i.e. with ε far away from the machine precision) and the spectral version are significantly lower than these usually committed by the latter in an operational context.

Then, the split-explicit schemes usually used in HEVI models, have been defined as a reference in terms of scalability in this study. Since the application of the operators controlling the horizontal propagation of the fastest waves in both HEVI and SI models are similar in terms of computation cost and exchange of data, the scalability study has been reduced to only one synthetic metric.

Moreover, experiments have been carried out to be informative whatever the future strategy planned for the time step: keeping large time steps so as to keep the ratio of spatial resolution to temporal resolution constant, as has

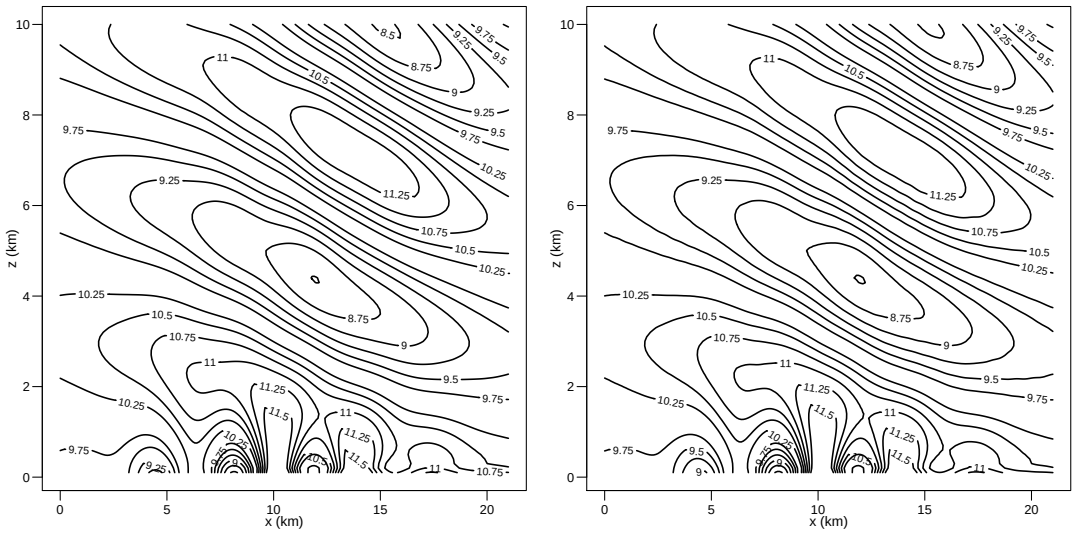

Figure 6. Mountain waves simulation with Δx = 100 m and Δt = 10 s, horizontal wind after 8000 s. Left: reference for the GP version; right: with a residual reduction ε^{-1} of 13500.

been done up to now, or gradually reducing this ratio to better represent some of the non-hydrostatic effects.

Whatever the strategy considered, we have shown that the grid-point version of the AROME model is at least as scalable as an HEVI model regardless of the time step chosen, thus confirming the interest of keeping the SI dynamics in addition to its main features for the future. Besides, although no scalability comparison has been made in this study between the spectral and grid-point versions, the inevitable overhead that occurs when the time step decreases at a given resolution turns out to be relatively lower in the grid-point version than in the spectral one. In doing so, by bypassing the expensive spectral transform steps, the grid-point method appears to be more promising than the spectral one if a time step reduction strategy is adopted.

These promising results have led to the implementation of an almost entirely grid-point version in the AROME model, with the exception of the diffusion, which remains performed implicitly in the spectral space for the time being. The first results obtained (about ten forecast days over one quarter of the current operational domain) confirm the main results presented in this study: the viability of keeping the main characteristics of the current operational model and, the fast convergence rate of the solver even with high-resolution real-case flows on the current operational vertical grid.

Furthermore, in view of using models with hectometric resolution on the horizontal in the next few years, thus better representing orography and in particular steeper slopes, it might be necessary to relax the assumption of flat terrain in the definition of the SI base state and to include some of the orographic terms in the implicit problem. Preliminary studies show that the stability of the SI scheme in the grid-point version presented here may be significantly improved by doing so.

A. Governing equations

The vertical coordinate η is related to hydrostatic pressure π by :

$$\pi(x,\eta,t) = A(\eta) + B(\eta)\,\pi_s(x,t), \tag{24}$$

where A and B are functions used in the definition of the vertical coordinate η, and $\pi_s(x,t)$ is the surface hydrostatic pressure (Laprise (1992)). Governing equations for the dry two-dimensional system are :

$$\frac{dU}{dt} + \frac{R_d T}{p}\frac{\partial p}{\partial x} + \frac{1}{m}\frac{\partial p}{\partial \eta}\frac{\partial \phi}{\partial x} = \nu,$$

$$\begin{aligned}\frac{dd}{dt} + g^2\frac{p}{mR_dT}\frac{\partial}{\partial \eta}\left(\frac{1}{m}\frac{\partial(p-\pi)}{\partial \eta}\right) - g\frac{p}{mR_dT}\frac{\partial U}{\partial \eta}\frac{\partial w}{\partial x}\\ -d\left[\frac{\partial U}{\partial x} - \left(\frac{\partial U}{\partial x} + d + \frac{p}{R_dTm}\frac{\partial \phi}{\partial x}\frac{\partial U}{\partial \eta}\right)\right]\\ = -g\frac{p}{mR_dT}\frac{\partial W}{\partial \eta},\end{aligned}$$

$$\frac{dT}{dt} + \frac{R_dT}{C_{vd}}\left(\frac{\partial U}{\partial x} + d + \frac{p}{R_dTm}\frac{\partial \phi}{\partial x}\frac{\partial U}{\partial \eta}\right) = \frac{Q}{C_{vd}},$$

$$\frac{d\hat{q}}{dt} + \frac{C_{pd}}{C_{vd}}\left(\frac{\partial U}{\partial x} + d + \frac{p}{R_dTm}\frac{\partial \phi}{\partial x}\frac{\partial U}{\partial \eta}\right) + \frac{\dot{\pi}}{\pi} = \frac{Q}{C_{vd}T},$$

$$\frac{\partial \pi_s}{\partial t} + \int_0^1 \frac{\partial (mU)}{\partial x}\,d\eta = 0.$$

The following diagnostic relations are:

$$m = \frac{\partial \pi}{\partial \eta},$$

$$p = \pi e^{\hat{q}},$$

$$\phi = \phi_s + \int_\eta^1 \frac{mRT}{p} d\eta,$$

$$\dot{\pi} = U\frac{\partial \pi}{\partial x} - \int_0^\eta \frac{\partial (mU)}{\partial x} d\eta',$$

$$g\frac{\partial w}{\partial x} = g\frac{\partial w_s}{\partial x} + \int_\eta^1 \frac{mR_dT}{p}\frac{\partial d}{\partial x} d\eta' + \int_\eta^1 R_d d \frac{\partial}{\partial x}\left(\frac{mT}{p}\right) d\eta'.$$

B. Notations used

We use the following notation:

- p : local pressure
- w : vertical velocity
- ϕ : geopotential
- g : acceleration of gravity
- R_d : gas constant of dry air
- C_p : specific heat capacity of dry air at constant pressure
- C_v : specific heat capacity of dry air at constant volume
- ν, W, Q : physical contributions for U, w and T respectively

The underscript 's' refers to a surface field.

C. Vertical operators of $\mathcal{L}$

$$\partial^* X = \frac{\pi^*}{m^*}\frac{\partial X}{\partial \eta}, \tag{25}$$

$$\mathcal{G}^* X = \int_\eta^1 \frac{m^*}{\pi^*} X d\eta, \tag{26}$$

$$\mathcal{S}^* = \frac{1}{\pi^*}\int_0^\eta m^* X d\eta, \tag{27}$$

$$\mathcal{N}^* X = \frac{1}{\pi_s^*}\int_0^1 m^* X d\eta, \tag{28}$$

$$\mathcal{L}^* X = \partial^*(\partial^* + 1)X, \tag{29}$$

where $\pi^*(\eta) = A(\eta) + B(\eta)\,\pi_s^*$ is the hydrostatic pressure of the basic state and $m^* = \partial \pi^*/\partial \eta$ is its vertical metric factor associated.

References

Auger, L., Dupont, O., Hagelin, S., Brousseau, P., and Brovelli, P. (2015). Arome–nwc: a new nowcasting tool based on an operational mesoscale forecasting system. Quarterly Journal of the Royal Meteorological Society, 141(690):1603–1611.

Axelsson, O. (1996). Iterative Solution Methods. Cambridge University Press.

Bauer, P., Thorpe, A., and Brunet, G. (2015). The quiet revolution of numerical weather prediction. Nature, 525(7567):47–55.

Bénard, P. (2003). Stability of semi-implicit and iterative centered-implicit time discretizations for various equation systems used in nwp. Monthly weather review, 131(10):2479–2491.

Bénard, P. and Glinton, M. R. (2019). Circumventing the pole problem of reduced lat–lon grids with local schemes. part i: Analysis and model formulation. Quarterly Journal of the Royal Meteorological Society, 145(721):1377–1391.

Bénard, P., Vivoda, J., Masek, J., Smolíková, P., Yessad, K., Smith, C., Brozková, R., and Geleyn, J.-F. (2010). Dynamical kernel of the aladin-NH spectral limited-area model: Revised formulation and sensitivity experiments. Quarterly Journal of the Royal Meteorological Society, 136(646):155–169.

Bryan, G. H. and Fritsch, J. M. (2002). A benchmark simulation for moist nonhydrostatic numerical models. Monthly Weather Review, 130(12):2917–2928.

Bubnová, R., Hello, G., Bénard, P., and Geleyn, J.-F. (1995). Integration of the fully elastic equations cast in the hydrostatic pressure terrain-following coordinate in the framework of the ARPEGE/aladin NWP system. Monthly Weather Review, 123(2):515–535.

Carpenter Jr, R. L., Droegemeier, K. K., Woodward, P. R., and Hane, C. E. (1990). Application of the piecewise parabolic method (ppm) to meteorological modeling. Monthly Weather Review, 118(3):586–612.

Durran, D. R. (2013). Numerical methods for wave equations in geophysical fluid dynamics, volume 32. Springer Science & Business Media.

Gassmann, A. (2013). A global hexagonal c-grid non-hydrostatic dynamical core (icon-iap) designed for energetic consistency. Quarterly Journal of the Royal Meteorological Society, 139(670):152–175.

Haugen, J. E. and Machenhauer, B. (1993). A spectral limited-area model formulation with time-dependent boundary conditions applied to the shallow-water equations. Monthly Weather Review, 121(9):2618–2630.

Hess, R. and Joppich, W. (1997). A comparison of parallel multigrid and a fast fourier transform algorithm for the solution of the helmholtz equation in numerical weather prediction. Parallel Computing, 22(11):1503–1512.

Klemp, J. B., Skamarock, W. C., and Dudhia, J. (2007). Conservative split-explicit time integration methods for the compressible nonhydrostatic equations. Monthly Weather Review, 135(8):2897–2913.

Kühnlein, C., Deconinck, W., Klein, R., Malardel, S., Piotrowski, Z. P., Smolarkiewicz, P. K., Szmelter, J., and Wedi, N. P. (2019). Fvm 1.0: a nonhydrostatic finite-volume dynamical core for the ifs. Geoscientific Model Development, 12(2):651–676.

Laprise, R. (1992). The euler equations of motion with hydrostatic pressure as an independent variable. Monthly weather review, 120(1):197–207.

Liesen, J. and Tichý, P. (2004). Convergence analysis of krylov subspace methods. GAMM-Mitteilungen, 27(2):153–173.

Mesinger, F. (1979). Dependence of vorticity analogue and the rossby wave phase speed on the choice of horizontal grid. Bulletin (Académie serbe des sciences et des

arts. Classe des sciences mathématiques et naturelles. Sciences mathématiques), pages 5–15.

Müller, E. H. and Scheichl, R. (2014). Massively parallel solvers for elliptic partial differential equations in numerical weather and climate prediction. Quarterly Journal of the Royal Meteorological Society, 140(685):2608–2624.

Pinty, J.-P., Benoit, R., Richard, E., and Laprise, R. (1995). Simple tests of a semi-implicit semi-lagrangian model on 2d mountain wave problems. Monthly Weather Review, 123(10):3042–3058.

Qaddouri, A. and Lee, V. (2010). The elliptic solvers in the canadian limited area forecasting model gem-lam. In Modeling Simulation and Optimization-Tolerance and Optimal Control. IntechOpen.

Saad, Y. and Schultz, M. H. (1986). Gmres: A generalized minimal residual algorithm for solving nonsymmetric linear systems. SIAM Journal on scientific and statistical computing, 7(3):856–869.

Satoh, M. (2002). Conservative scheme for the compressible nonhydrostatic models with the horizontally explicit and vertically implicit time integration scheme. Monthly Weather Review, 130(5):1227–1245.

Schär, C., Leuenberger, D., Fuhrer, O., Lüthi, D., and Girard, C. (2002). A new terrain-following vertical coordinate formulation for atmospheric prediction models. Monthly Weather Review, 130(10):2459–2480.

Seity, Y., Brousseau, P., Malardel, S., Hello, G., Bénard, P., Bouttier, F., Lac, C., and Masson, V. (2011). The AROME-France convective-scale operational model. Monthly Weather Review, 139:976–991.

Simmons, A. J., Hoskins, B. J., and Burridge, D. M. (1978). Stability of the semi-implicit method of time integration. Monthly Weather Review, 106(3):405–412.

Skamarock, W. C., Smolarkiewicz, P. K., and Klemp, J. B. (1997). Preconditioned conjugate-residual solvers for helmholtz equations in nonhydrostatic models. Monthly weather review, 125(4):587–599.

Smolarkiewicz, P. and Margolin, L. (1994). Variational solver for elliptic problems in atmospheric flows. Appl. Math. Comp. Sci, 4(4):527–551.

Soleymani, F. (2013). A new method for solving ill-conditioned linear systems. Opuscula Mathematica, 33.

Steppeler, J., Hess, R., Schättler, U., and Bonaventura, L. (2003). Review of numerical methods for nonhydrostatic weather prediction models. Meteorology and Atmospheric Physics, 82(1):287–301.

Thomas, S. J., Malevsky, A. V., Desgagné, M., Benoit, R., Pellerin, P., and Valin, M. (1997). Massively parallel implementation of the mesoscale compressible community model. Parallel Computing, 23(14):2143–2160.

Wedi, N. P. (2014). Increasing horizontal resolution in numerical weather prediction and climate simulations: illusion or panacea? Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 372(2018):20130289.

Wicker, L. J. and Skamarock, W. C. (1998). A time-splitting scheme for the elastic equations incorporating second-order runge–kutta time differencing. Monthly Weather Review, 126(7):1992–1999.

Ye, Q. (2017). Preconditioning for accurate solutions of linear systems and eigenvalue problems. arXiv preprint arXiv:1705.04340.

Zheng, Y. and Marguinaud, P. (2018). Simulation of the performance and scalability of message passing interface (mpi) communications of atmospheric models running on exascale supercomputers. Geoscientific Model Development, 11(8):3409–3426.