Development of the global multiscale atmosphere model: computational aspects

M A Tolstykh\textsuperscript{1,2,3}, G S Goyman\textsuperscript{1,3}, R Y Fadeev\textsuperscript{1,3}, S V Travova\textsuperscript{2} and V V Shashkin\textsuperscript{1,2,3}

\textsuperscript{1}Marchuk Institute of Numerical Mathematics Russian Academy of Sciences, Moscow, Russia
\textsuperscript{2}Hydrometcentre of Russia, Moscow, Russia
\textsuperscript{3}Moscow Institute of Physics and Technology, Dolgoprudny, Russia

m.tolstykh@inm.ras.ru, gordeygoyman@gmail.com

Abstract. The global atmosphere model SL-AV is applied for operational numerical weather prediction in Russia at the time scales from days to months and also for simulation of modern climate. It is important to achieve the best wall-clock time for all of the model applications with different grid sizes. In this paper, we analyse the effect of switching computations and parallel exchanges from double to single precision in the semi-Lagrangian advection and elliptic equations solver blocks on the quality of the medium-range forecast and evaluate the change in wall-clock time. We have also optimized OpenMP parallelization in the computations of the right-hand sides of meteorological prognostic equations. The effect of these changes is studied for SL-AV model versions with different resolutions.

1. Introduction
The global atmosphere model is a powerful tool applied in a range of applications. These applications are the numerical weather prediction at time scales from hours to months, climate outlook for next few years and also climate change projections. All these applications require large computational resources because an atmosphere model solves numerically the system of 3D nonstationary simplified Navier-Stokes equations with quite cumbersome right hand sides describing the adiabatic processes that cannot be represented on a given finite grid (e.g. deep convection producing heavy rain, cloudiness, solar radiation, planetary boundary layer). In addition to the complexity of the numerical integration of governing equations, each of the abovementioned applications has its own peculiarities that impose additional requirements on computing resources: operational wall-clock time limit for the medium-range weather forecasting (usually, from 6 to 15 minutes per forecast day), ‘model years per day’ performance metric for climate modeling and necessity of many model runs at a time in case of ensemble forecasting. At the same time, the increase in grid resolution is one of the common ways to reduce model errors. So the computational and parallel efficiency is crucial for the atmosphere model
code at every resolution, and every weather prediction centre makes the tremendous efforts to make their model run faster on modern massively parallel computer systems.

In this paper, we describe our recent work to accelerate the parallel code of Russian global atmosphere model SL-AV (Semi-Lagrangian, based on Absolute Vorticity equation) [1], applied for operational numerical weather prediction at different time scales and for climate modelling. We consider two ways to speed up calculations – partial use of single precision floating point arithmetic for calculations and parallel communications in the semi-Lagrangian advection and elliptic equations solver blocks, and also optimization of the use of cache memory in the block for calculating the right-hand sides of the equations due to parameterizations of subgrid scale processes. The particular feature of this study is that we are interested in code acceleration for every model resolution that we apply, ranging from the climate change projection (grid dimension of $10^9$) to the high-resolution medium-range weather forecast where the grid dimension tends to $10^{10}$. Besides the code acceleration issues, we consider the use of the single precision as a promising approach to improve the parallel efficiency of the elliptic solver and semi-Lagrangian algorithms implementations since it allows to decrease the amount of data in interprocessor exchanges.

The paper is organized as follows: Section 2 presents the configurations of the considered versions of the SL-AV model. Section 3 contains a description of the work on optimizing memory use in the subgrid scale parameterizations block. Section 4 outlines the implementation of the single precision. The final profiling of SL-AV model as well as conclusions are presented in Section 5.

2. SL-AV model and parallel system configurations

SL-AV is the global atmospheric model developed at the Marchuk Institute of Numerical Mathematics Russian Academy of Sciences and Hydrometcentre of Russia [1]. It consists of original dynamical core [2] and parameterizations of subgrid-scale processes mostly developed by ALADIN/LACE consortium [3]. The calculations of the short-wave and long-wave atmosphere radiation use known software packages CLIRAD-SW [4, 5] and RRTMG LW [6]. The hydrostatic dynamical core is based on the semi-implicit semi-Lagrangian approach used in many operational numerical weather prediction models [7]. Its distinct features are the use of the vertical component of the absolute vorticity and horizontal divergence as prognostic variables, unstaggered grid, and fourth-order finite differences.

There are several versions of the model based on the same code. There is a version for operational medium-range weather forecasts with the horizontal resolution of 0.225 degrees in longitude and variable resolution in latitude from 0.24 degrees in Southern hemisphere to 0.16 in Northern one. This is about 20 km in Northern mid-latitudes. The model vertical grid has 51 levels. The version for ensemble prediction system has the horizontal resolution of 0.72x0.9 degrees in latitude and longitude respectively, and 96 vertical levels. This gives approximately 75 km resolution in mid-latitudes. This version is also used in the Hydrometcentre of Russia long-range probabilistic forecast system and for modeling modern climate state [8, 9]. Further, we call them “20 km” and “70 km” version, respectively. We are also developing a new version of the SL-AV model that has the horizontal resolution of ~11 km and 104 vertical levels. Some preliminary results of parallel code efficiency for this resolution are published in [10, 11].

Numerical experiments presented in this study are performed at Cray XC40 supercomputer installed at Roshydromet’s Main Computing Center. This supercomputer consists of 936 nodes with Intel Xeon E2697v4 18-core CPUs and 256 GB memory, connected via Cray ARIES interconnect.

We use eight nodes with 8 MPI-processes per node and 4 OpenMP threads per MPI-process to run model version with 70 km horizontal resolution. 27 nodes with the same per node MPI/OpenMP configuration are applied for running the 20 km model version. These per node configurations of MPI and OpenMP are typical for these versions of the SL-AV model at the respective computer systems and found to be nearly optimal.
3. Memory usage optimization

The so-called physical parameterizations are used in atmospheric models to account for the impact of physical processes that are not resolvable at the computational grid. The parameterizations play a crucial role in the accuracy of atmospheric modeling, being one of the most computationally expensive parts of the model.

The implementation details of the physics parameterizations block in the SL-AV model are described in [12]. We use OpenMP parallelization in this block. In the original version of the SL-AV model, we use the approach based on the additional decomposition of the computational domain into a set of 2d subdomains of the size \( \frac{NLON}{NOMP} \times NLEV \), where NLON is the grid dimension in longitude, NOMP is the number of OpenMP threads and NLEV is the dimension of the vertical grid. The number of such subdomains are equal to the number of latitudes corresponding to a given MPI-process. These subdomains are distributed among the OpenMP threads and each OpenMP thread performs all computations in the block of physical parameterizations sequentially for each subdomain. A schematic representation of the domain decomposition and subdomains distribution between OpenMP threads is shown in figure 1.

![Schematic of the approach used in the SL-AV model for partitioning the computational domain among OpenMP threads](image)

**Figure 1.** Schematics of the approach used in the SL-AV model for partitioning the computational domain among OpenMP threads. Each block represents a 2d array with a size \( \frac{NLON}{NOMP} \times NLEV \).

In our previous work [11], we have tested a possibility of additional partitioning of the computational domain for each OpenMP thread. The idea behind this approach is to reduce the size of the arrays being involved in the computations thereby improving data locality, reducing the number of cache misses and increasing the efficiency of using the processors’ memory. The tests with the prototype version of the SLAV model with 10 km horizontal resolution have shown that it is possible to reduce elapsed time of the parameterization block by 26% varying the size of the subdomains at each OpenMP thread. In this work, we have implemented a more efficient version of this approach.

Now, instead of splitting the computational domain at each OpenMP thread, we perform the decomposition of the entire computational domain into blocks of the size \( \frac{NLON}{NB} \times NLEV \), and redistribute these blocks across OpenMP threads. In this case, there are \( NB \times NJ \) blocks at each MPI-process, where NJ is the number of grid points along the latitude corresponding to a given MPI-process (1D MPI-decomposition along latitudinal direction is used in the SL-AV model). A schematic representation of this approach is shown in figure 2. Application of the new approach allows to select the optimal size of computational blocks regardless of OpenMP threads number and the horizontal resolution of the computational grid being used.
Figure 2. Schematics of the new approach for partitioning the computational domain among OpenMP threads. Each block represents a 2d array with the size \( \frac{N_{LON}}{NB} \times N_{LEV} \).

Figure 3 shows the relative change in the average elapsed time of the computations in the block of parameterizations for 20 km and 70 km versions of the SL-AV model after the implementation of additional domain partitioning. For these versions of the model, a decrease in computations time by 37% and 14% respectively is observed. The optimal block size in both cases turned out to be \( \frac{N_{LON}}{NB} = 16 \) that corresponds to 2-4 vector register lengths for double precision floating point numbers, depending on the used set of vector instructions. The greater performance gain for the 20 km version of the model is apparently explained by the smaller vertical grid size that leads to a smaller computational block size.

![Physics parametrizations elapsed time](image)

**Figure 3.** Relative reduction of physics parametrizations elapsed time for the 20 km and 70 km resolution of SL-AV model. The bars show time reduction relative to the reference version of the code.

### 4. Implementing single precision

There are always uncertainties in numerical weather forecasting and climate modeling associated with the inaccuracy of determining the initial state of the atmosphere, errors in methods of approximating the equations of atmospheric dynamics (this includes both the errors in calculating differential operators and the errors in parameterizations of subgrid scale processes), rounding and numbers representation errors of floating point arithmetic. It can be assumed that in some cases round-off errors make an insignificant contribution to the overall uncertainty in comparison with other sources. At the
same time, the selective use of single precision instead of double precision commonly used in numerical weather prediction for computations and parallel communications can significantly reduce the required computer resources. This is important as the numerical weather prediction is a time-critical application requiring thousands of processor cores.

To date, there is a number of theoretical [13, 14] and practical studies [15, 16, 17, 18] devoted to the possibility of using reduced accuracy for atmospheric modeling. For instance, the experience of the European Center for Medium-Range Weather Forecasts with the Integrated Forecast System [16] shows that using single precision can save up to 40% of computational time without noticeably affecting the accuracy of a medium-range ensemble weather forecast. Thus, we consider the use of single precision as one of possible approaches to increase the computational and parallel efficiency of the SL-AV model.

In [11], we have implemented a possibility of single precision usage to implement parallel data transposition in the block for solving elliptic equations as well as parallel communications and computations in the semi-Lagrangian advection block. These parts of the model are among the most time consuming, and also are the most intensive in terms of parallel communications.

Tests with a prototype of a new 10 km SL-AV model version have shown that the implementation of single precision in these blocks can reduce the execution time of the semi-Lagrangian advection block by 24% and the execution time of the elliptic solver by 22% while using 3888 cores of Cray-XC40 supercomputer. At the same time, the influence of the use of single precision on the weather forecast quality has not been studied in detail, since this version of the model is under development. In this paper, we investigate this issue within the 20 km version of SL-AV model used for operational medium-range weather forecasting at the Hydrometeorological Center of Russia.

To evaluate the changes in forecast accuracy, we have computed 31 120-hours forecasts from initial data of 00 UTC (universal coordinated time, coincides with Greenwich mean time) of each day of January 2020. The standard forecast error measures are calculated following the World meteorological organization standard [19]. These metrics include mean error (bias) and root-mean-squared error. All the metrics for all the standard regions are calculated, along with respective confidence intervals. Figures 4 and 5 show metrics for most deviating scores, while other regions and metrics show less differences. One can see that the introduction of single precision computations into the semi-Lagrangian part and data transpositions does not change the forecast error metrics within statistical significance.

![Figure 4. Root-mean-squared error (RMSEA) and mean error (RCOA) for wind vector at 250 hPa surface in Asia region. Black lines in the middle of each column show confidence intervals.](image)
Figure 5. Root-mean-squared error (RMS) and mean error (RCO) the height of 500 hPa surface in the tropics (20S-20N).

We have also measured the change in the elapsed time of these blocks. Figure 6 shows the relative change in average execution time in the semi-Lagrangian advection block and the block for solving elliptic equations for the 20 km and 70 km versions. One can see that the use of single precision can reduce the execution time of these blocks by 30% for both versions of the model. At the same time, the execution time of parallel exchanges in the semi-Lagrangian advection block and elliptic solver block was approximately halved, which is important from the point of view of model scalability since the phase of parallel exchanges in these blocks is one of the main bottlenecks in the SLAV model in terms of parallel efficiency [10].

Figure 6. Semi-Lagrangian advection and elliptic solver parts of the code elapsed time for the 20 km and 70 km resolution SL-AV model before and after introduction of single precision. The figures above the bars show the time reduction relative to the reference version of the code.
Figure 7. Percentage of the time used in different parts of the 20km and 70km version of SL-AV model before and after the optimizations. The figures inside the bars show the time reduction relative to the reference version of the model.

5. Conclusion
The global atmosphere model SL-AV is applied for a range of applications, from operational high-resolution forecasts to modern climate modelling. All these applications are quite computationally expensive. We consider two ways to accelerate the code of the global atmosphere model SL-AV with the horizontal resolution of 70 and 20 km. The first method is to optimize the use of the processor's memory access in the block for calculating the parameterizations of subgrid scale physical processes. To do this, we have implemented the possibility to partition the computational domain in this block into smaller subregions, so all the computations in the parametrizations block are carried out for such subregion at a time. The proposed approach requires minimal modification of the model code and makes it possible to reduce the elapsed time of parametrizations block by 14.5% for the SL-AV model with the horizontal resolution of 70 km and by 37.3% for a model with the resolution of 20 km. The reduction of the wall-clock time for the whole model is 8.2% and 17.9%, respectively (see figure 7).

The second way is to implement the use of single precision floating point arithmetic to perform parallel exchanges and part of the computations in the semi-Lagrangian advection and elliptic equations solvers blocks. An important issue here is the impact of using single precision on the forecast quality. To study this issue, we calculated a series of forecasts with the lead time of 120 hours. The results of these forecasts show the absence of statistically significant changes in the error metrics. At the same time, the introduction of single precision computations allows us to reduce the elapsed time of the semi-Lagrangian advection block approximately by 30%. This leads to the speedup of 6.5% and 5.2% of the whole model with the resolution of 20 km and 70 km respectively (see figure 7). We expect these changes to increase the parallel efficiency of the model since semi-Lagrangian advection and elliptic equations solvers are among the bottlenecks of the atmospheric model codes. Thus, we find it promising to further implement the use of single precision in the SL-AV model. We plan to investigate model sensitivity to the use of single precision within the framework of ensemble forecasting and climate modeling and to extend this approach to other blocks of the model and to the multigrid version [12] of the elliptic solver algorithm.

Acknowledgments
Computations presented in this article were carried out at the Roshydromet Main Computer Center (MCC). This study was carried out at Marchuk Institute of Numerical Mathematics, Russian Academy of Sciences and supported by Russian Foundation for Basic Researches grant No. 19-31-90032. The work described in Section 3 was funded by Moscow Center for Fundamental and Applied
Mathematics (agreement with the Ministry of Education and Science of the Russian Federation No. 075-15-2019-1624).

References
[1] Tolstykh M A, Fadeev R Y, Shashkin V V, Goyman G S, Zaripov R B, Kiktev D B, Makhnorylova S V, Mizyak V G and Rogutov V S 2018 Multiscale global atmosphere model SL-AV: the results of medium-range weather forecasts. Russ. Meteorol. Hydrol. 43 773–79
[2] Tolstykh M, Shashkin V, Fadeev R and Goyman G 2017 Vorticity-divergence semi-Lagrangian global atmospheric model SL-AV20: dynamical core Geosci. Model Dev. 10 1961–83
[3] Termonia P, Fischer C, Bazile E, Bouyssel F, Brozková R, Bénard P, Bochenek B, Degraauwe D, Derková M, Khatib R E et al. 2018 The ALADIN System and its canonical model configurations AROME CY41T1 and ALARO CY40T1 Geosci. Mod. Devel. 11 257–81
[4] Chou M D and Suarez M J 1999 A solar radiation parameterization (CLIRAD-SW) for atmospheric studies NASA Tech. Memo. 10460 48 pp
[5] Tarasova T and Fomin B 2007 The use of new parameterizations for gaseous absorption in the CLIRAD-SW solar radiation code for models J. Atmos. and Oc. Technology 24 1157–62
[6] Mlawer E J, Taubman S J and Brown P D 1997 RRTM, a validated correlated-k model for the longwave J. Geophys. Res. 102 16663–82
[7] Staniforth A and Cote J 1991 Semi-Lagrangian integration schemes for atmospheric models – a review Mon. Weather Rev. 119 2206–33
[8] Fadeev R Y, Tolstykh M A and Volodin E M 2019 Climate version of the SL-AV global atmospheric model: development and preliminary results Russ. Meteorol. and Hydrol. 44 37–8
[9] Fadeev R Y, Ushakov K V, Kalmykov V V, Tolstykh M A and Ibraye R A 2016 Coupled atmosphere–ocean model SLAV–INMIO: implementation and first results Russ. J. of Num. An. and Math. Mod., 31 329–37
[10] Tolstykh M, Goyman G, Fadeev R, Shashkin V and Lubov S 2019 SL-AV model: numerical weather prediction at extra-massively parallel supercomputer RuSCDays 2018. Communications in Computer and Information Science vol 965 ed Voevodin V, Sobolev S (Cham: Springer) pp 379–87
[11] Tolstykh M, Goyman G, Fadeev R and Shashkin V 2020 Implementation of SL-AV global atmosphere model with 10km horizontal resolution Communications in Computer and Information Science RuSCDays 2020 vol 1331 ed Voevodin V, Sobolev S (Springer) (in press)
[12] Tolstykh M, Goyman G, Fadeev R and Shashkin V 2018 Structure and algorithms of SL-AV atmosphere model parallel program complex Lobachevskii J. of Math. 39 587–95
[13] Thorners T, Dübner P D and Palmer T N 2017 On the use of scale-dependent precision in earth system modelling Quart. J. Roy. Meteor. Soc. 143 897–908
[14] Palmer T N 2014 More reliable forecasts with less precise computations: a fast-track route to cloud-resolved weather and climate simulators? Philos. Trans. Roy. Soc. A 372 20130391
[15] Baede A, Dent D and Hollingsworth A 1976 The effect of arithmetic precision on some meteorological integrations Technical Report 2 ECMWF [Available online at http://www.ecmwf.int/en/elibrary/7869-effect-arithmetic-precisionssome-meteorological-integrations.]
[16] Váňa F, Düben P, Lang S, Palmer T, Leutbecher M, Salmond D and Carver G 2017 Single precision in weather forecasting models: an evaluation with the IFS Mon. Wea. Rev. 145 495–502
[17] Nakano M, Yashiro H, Kodama C and Tomita H 2018 Single precision in the dynamical core of a nonhydrostatic global atmospheric model: evaluation using a baroclinic wave test case
[18] Russell F P, Niu X, Luk W and Palmer T N 2015 On the use of programmable hardware and reduced numerical precision in earth-system modeling *J. Adv. Model. Earth Syst.* 7 1393–1408

[19] World Meteorological Organization 2017 *Manual on the Global Data-processing and Forecasting System* (Geneva: WMO No. 485)