ECMWF Feature article

from Newsletter Number 148 – Summer 2016

METEOROLOGY

Single-precision IFS

Image from Mallivan/iStock/Thinkstock

www.ecmwf.int/en/about/news-centre/media-resources

doi:10.21957/avq2i3tp

This article appeared in the Meteorology section of ECMWF Newsletter No. 148 – Summer 2016, pp. 20-23.

Single-precision IFS

Filip Váňa, Glenn Carver, Simon Lang, Martin Leutbecher, Deborah Salmond (all ECMWF) Peter Düben, Tim Palmer (both University of Oxford)

Since the early days of numerical weather prediction (NWP), the issue of appropriate numerical precision has been the subject of considerable interest. Indeed, ECMWF's second Technical Report, published in 1976 by Baede at al., was devoted to 'The effect of arithmetic precision on some meteorological integrations'. Some precision-sensitive operations, such as matrix inversion, often require at least so-called double-precision arithmetic (i.e. a 64-bit representation of real numbers) to deliver acceptable results. With the growth of computer power and especially the availability of 64-bit processors, double precision came to be commonly used for all floating-point computations in numerical modelling, including in NWP.

However, such precision may be wasteful since it uses up precious computing resources while not necessarily making much difference to forecast quality. Building on work carried out at the University of Oxford, we have made single-precision arithmetic (a 32-bit representation of real numbers) available for ensemble forecasts in ECMWF's Integrated Forecasting System (IFS). Experiments show that, at a horizontal resolution of 50 km, single precision brings significant savings in computational cost without degrading forecast quality.

Computing challenges

Numerical weather prediction relies on powerful supercomputers to carry out the complex calculations on which forecasts are based. Limitations in the available computing power represent a significant obstacle to increasing the accuracy of such forecasts. In current architectures, the limiting factor is not so much peak processor performance but the speed of memory access.

As a result, there is growing interest in reducing the data volume being processed to the minimum information necessary to deliver high-quality forecasts. Intuitively, evaluating temperature to 16 valid digits might seem excessive given that in current NWP systems even the third digit is subject to significant error. Thus the important practical question is how much information needs to be included when weather or climate models are run on supercomputers. Assuming that only a fraction of the information contained in double-precision data is relevant, the key question is how this can be exploited in running the model: how can the redundant overhead be eliminated in order to profit computationally while still producing forecasts of equal skill? This question needs to be addressed and answered before possibly adapting the code to future computer hardware that offers a choice of floating-point precision or even variable levels of precision (see *Düben et al.*, 2015).

Considering possible future high-performance computing (HPC) architectures, there is also an exciting area of research into the development of approximate and stochastic computing hardware that allows a trade-off between precision and energy consumption. With this in mind, it is important to better understand the minimum computational precision requirements essential for successful weather forecasts.

Adapting the code

As a part of their involvement in the design of a future probabilistic Earth system model for climate prediction, a research group at the University of Oxford led by ECMWF Fellow Professor Tim Palmer has adapted the OpenIFS model to use single-precision arithmetic. The OpenIFS model is a version of the IFS which is available to universities under licence and in which the data assimilation code has been removed. It is ideally suited to this kind of proof-of-concept study because it offers a portable and much-reduced code whilst retaining all the forecast capability and code of the operational IFS. This pioneering exercise delivered very sensible forecasts while requiring only a few scientific code modifications.

The encouraging results have prompted ECMWF to make this option available in the IFS for ensemble forecasts. Since the IFS code is used in different configurations, implementing the single-precision option requires making both levels of precision available within the same source code. Single precision has been available as an alternative to the default double-precision arithmetic in IFS forecasts from model cycle 41r2, the current operational cycle at ECMWF. This article summarises the main results obtained with the single-precision IFS code.

Technical implementation

The key element defining model precision is the setting of the Fortran KIND parameter of all real variables in the IFS code. This means the precision is determined at the compilation stage and cannot be changed during the execution. Additional code changes were required to allow both single and double-precision functionality within the same code, mostly related to hard-coded numerical thresholds introduced to suit exclusively double precision; extending code interfaces with raw (i.e. binary) data; the MPI (Message Passing Interface) library interface; system functions; and mathematical libraries. Systematic experimentation with single-precision arithmetic revealed some poorly conditioned code, for which a more robust but scientifically equivalent formulation was sought to deliver equal performance for both precisions. Such changes were very beneficial to the overall code robustness regardless of the precision used.

Finally, a few areas of the code that were sensitive to precision had to continue to run strictly using double-precision arithmetic. As well as some specific double-precision computations in surface-scheme and shallow-convection physics, double precision must mainly be used in the setup part of the Galerkin methods used in the IFS: calculations of the roots and the associated polynomials for Legendre transformation, and the evaluation of integral operators for the vertical finite element (VFE) discretisation scheme. In both cases, double precision is only required for the (once-only) initial pre-computation of operators. The results are then truncated to the chosen precision used consistently for all other computations. These changes have been found to deliver equally skilled forecasts in both precisions up to TL399, which corresponds to about 50 km horizontal resolution, the highest resolution tested to date.

Forecast skill

To evaluate the skill of the single-precision IFS, two sets of experiments were performed: a long-range integration to test the model's ability to deliver a reasonable climate, and the production of medium-range ensemble forecasts (ENS) to compare the skill of single-precision ENS with that of reference double-precision ENS.

Long-term simulations

In order to assess the model climate and compare it against observations, a four-member ensemble was integrated for 13 months at TL399 resolution with 137 vertical levels. All integrations were run in uncoupled mode, with prescribed sea-surface temperatures, although coupling to the operational wave model was included. The first month of the integrations was discarded. The start of each member of the ensemble was shifted by 1 day and 6 hours in order to properly sample the diurnal cycle. The 12-month integration covering the period of September 2000 to August 2001 was compared with various datasets, including the ERA-Interim 60-level climatology. The model was configured in exactly the same way as the operational forecast model, the only difference being the activation of the pressure mass fixer to ensure mass conservation, which is necessary for long-term integrations.

The differences between the single-precision experiment and the double-precision reference run were generally small relative to the magnitude of systematic forecast errors. The main differences were observed in the mid-latitudes and the polar areas, typically over continental land masses. This can be attributed to the high flow variability and associated uncertainty of long-term integrations in these parts of the globe. The maximum difference of annual zonal averages between the two sets of climate simulations is less than 2% in relative humidity, 0.5 K in temperature and 1 m/s in wind speed. Compared to observations, the experiments performed equally well. This is illustrated by Figure 1, which shows the difference in top-of-atmosphere shortwave radiation flux between the annual climatology computed by each of the model versions on the one hand and data from the CERES EBAF satellite product on the other.

Medium-range ensemble forecasts

To investigate the single-precision IFS capability for ensemble forecasting, 46 single-precision and double-precision ensemble forecasts with 50 members each were run evenly distributed over an entire year, between 4 December 2013 and 29 November 2014. The resolution used was again TL399, this time with 91 vertical levels, as used for operational ensemble forecasts. The forecast range for each ensemble was 15 days.

In the extra-tropics, the results of the single and double-precision experiments are very similar for the first 12 days of the forecasts. After day 12, there are some small differences between the two experiments, e.g. for geopotential at 500 hPa (not shown). This could be an effect of the limited sample size. For temperature at 850 hPa, the results are very similar throughout the forecast period.

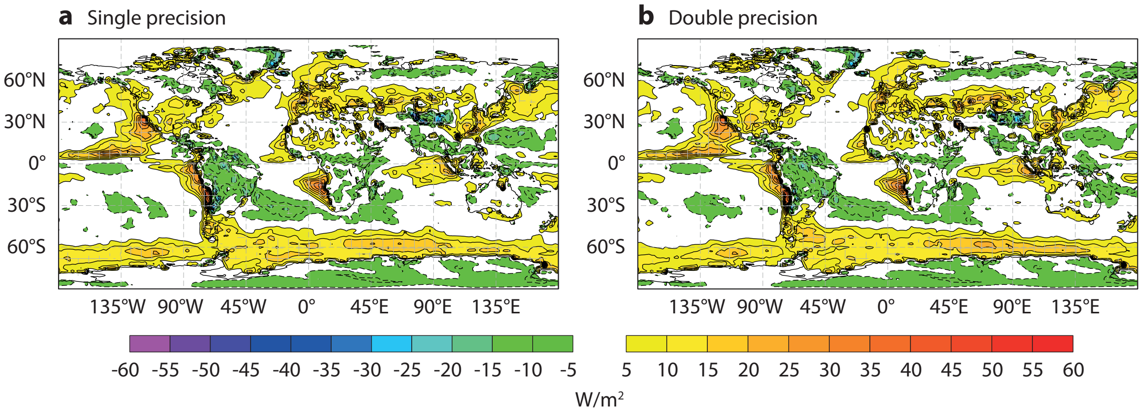

Figure 1 Difference between the mean top-of-atmosphere shortwave radiation flux in a 1-year IFS simulation at the resolution TL399 and in the CERES EBAF satellite data product for (a) the single-precision IFS and (b) the double-precision IFS. Hatching indicates statistically significant differences at the 95% confidence level.

In the tropics, a more striking difference between the experiments is observed for geopotential at 500 hPa: the single-precision forecasts systematically outperform the double-precision reference forecasts (Figure 2a). This difference stems from the non-conservation of mass in the IFS. The slight gain or loss of mass contributes to a warm or cold temperature bias in the tropics. It appears that the slightly less conservative single-precision version, for this particular combination of time step and resolution, compensates better for the existing model bias. When the mass fixer is activated for both experiments, the results are indistinguishable in the tropics too (Figure 2b–c).

Despite being slightly less conservative, the single-precision IFS delivered ensemble forecasts of equivalent skill compared to the double-precision reference forecasts.

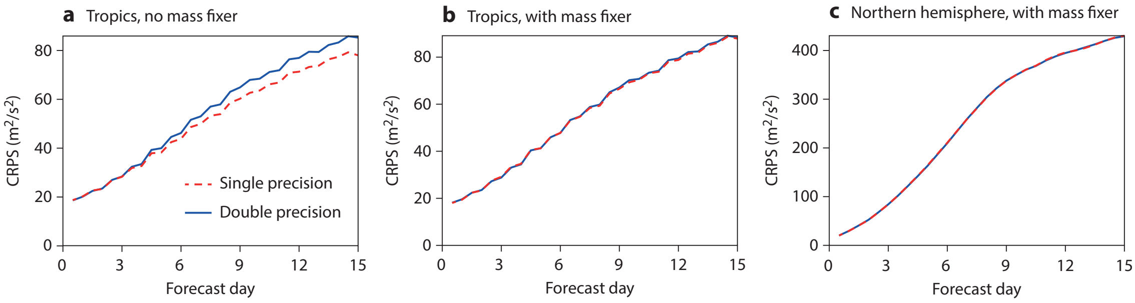

Figure 2 Continuous Ranked Probability Score (CRPS) for 15-day ensemble forecasts of geopotential at 500 hPa (a) in the tropics without mass fixer, (b) in the tropics with mass fixer and (c) in the northern hemisphere with mass fixer.

Computational cost

To evaluate the gain in computing performance brought by the single-precision IFS, the computational cost of the two sets of ensemble forecasts described in the previous section was compared on ECMWF's Cray XC30 HPC facility. All experiments were submitted with the same topology: 96 MPI tasks and eight OpenMP threads.

The reduction in computational cost per standard model time step when using the single-precision version was found to be 37%. When the difference in cache utilisation between the two versions is compensated for by doubling the length of the inner 'DO-loops' in the model (by adjusting the tunable namelist parameter NPROMA), this gain increased to 40.7%.

Figure 3 shows a breakdown of computational cost by model component. The results obtained for single and double precision with the same NPROMA parameter are surprisingly similar. This implies that all parts of the model benefit from single precision. However, the gain in the Fast Fourier Transformation code (FFT) is only 20%, which is significantly less than the overall gain of 37%. This probably confirms that the FFT code is already very efficient and that it is hard to speed it up any more. Another notable result is the 62% performance gain in grid-point dynamics using single precision. This shows that the current value of the NPROMA parameter (NPROMA = 16), which has been optimised for the model as a whole, is too big for this particular part of the IFS. The reduction in data size with single precision and the reduced memory access also result in better cache utilisation for this code part. Significant differences in relative computational cost can also be seen in the model components labelled ‘Other’ in Figure 3, which, among other things, include input/output (I/O) and setup. Here some parts have to be computed exclusively using double precision, and the precision of I/O is prescribed by GRIB packing. As a result, the relative cost of the ‘Other’ category increases in the single-precision IFS.

When doubling the NPROMA parameter in the single-precision run (to return to roughly the original cache utilisation), the breakdown diagram looks different (Figure 3c). Not surprisingly, the grid-point dynamics reverts to the original, less optimised performance. But this is the only component for which the absolute cost has increased compared to running the single version with NPROMA = 16. The other model parts are all improved, but their absolute performance gain is proportional to the ratio between computation and data manipulation. Notably the IFS physics, which involves a lot of computation with a relatively limited amount of data manipulation, benefits least from this final optimisation step.

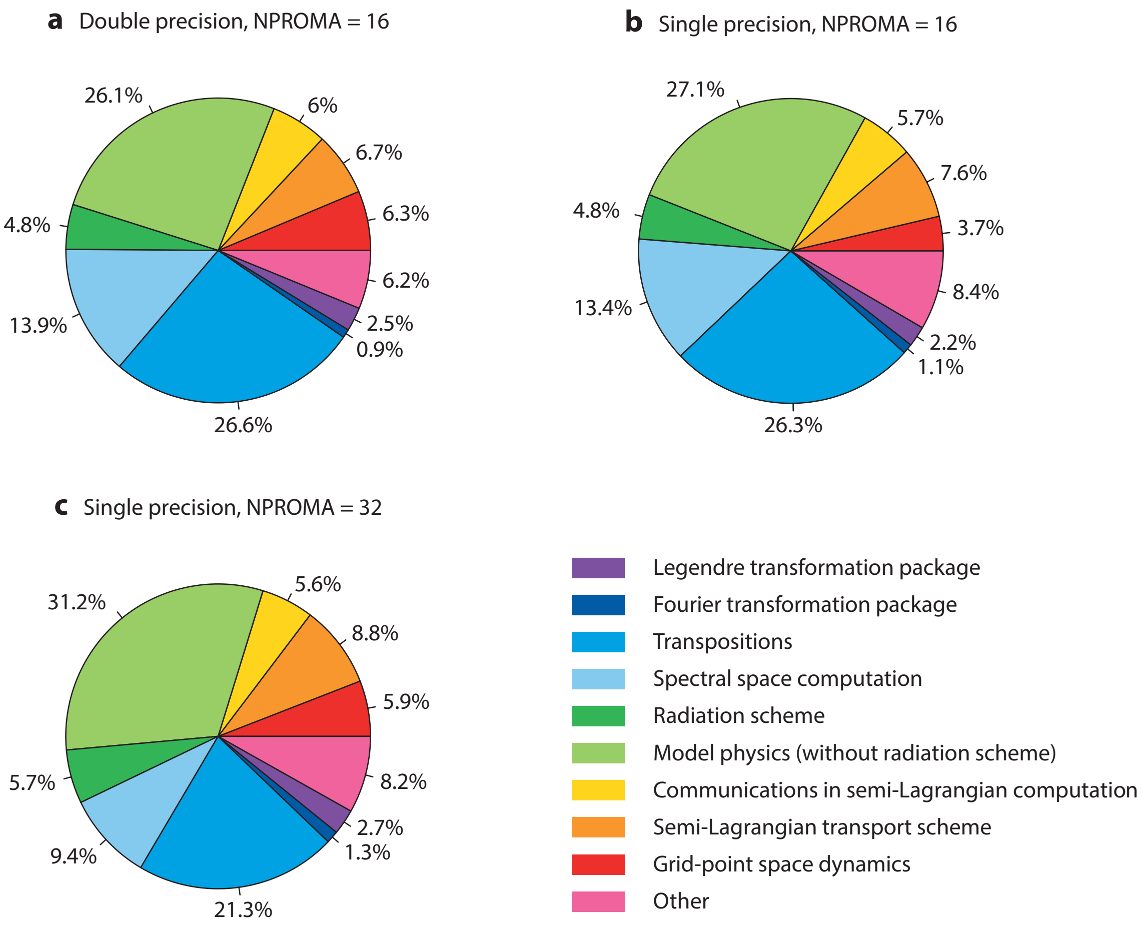

Figure 3 Breakdown of computational cost by model component shown for (a) the double-precision IFS with the default setting NPROMA = 16, (b) the single-precision IFS with NPROMA = 16 and (c) the single-precision IFS with NPROMA = 32. The results were obtained at the resolution TL399 with 91 vertical levels.

Discussion and outlook

A study using ECMWF’s OpenIFS model demonstrated the potential computational benefits of using reduced numerical precision in the production of ensemble forecasts. The single-precision arithmetic has now been successfully implemented in the IFS. According to first tests with ensemble forecasts at a horizontal resolution of 50 km, it offers almost indistinguishable forecast quality at about 40% greater computational efficiency. The savings in computational cost come from reduced memory access on ECMWF’s Cray XC30 high-performance computing facility. To achieve this, it was necessary to review the entire IFS code and to ensure that all components of the model are still delivering expected results at this precision. Overall, the results of the experiments demonstrate that there is a great potential to reduce power consumption and save computational time by investing in more sophisticated code that can handle reduced-precision arithmetic.

In addition to the reduced computational cost, the process of testing single precision using the IFS code has helped to detect and fix badly conditioned code. In the future, we would like to review more configurations with the single-precision alternative (such as the tangent-linear code) to reveal potentially problematic code.

The relatively straightforward success of applying single-precision arithmetic indicates that there is further potential for improved efficiency with future computing architectures. The work described here can serve as the starting point for further research into the use of reduced precision beyond single precision, and in particular into a flexible reduction of numerical precision depending on spatial scale, as proposed by Düben et al. (2014) and Düben et al. (2015). Future work at ECMWF will focus on exploring the impact of reduced precision at higher resolutions. As operational ensemble forecasts are produced in coupled mode with the NEMO ocean model, it is also desirable to extend the reduced-precision option from the atmospheric component to the whole coupled system. The single-precision option could also be explored for different configurations of the IFS, such as those used in data assimilation, observation processing and atmospheric composition. Finally, the ability to run the IFS in single precision will increase flexibility for ECMWF's Scalability Programme, which aims to prepare forecasting systems for the exascale era of supercomputing, for example when running the IFS on alternative hardware (such as GPUs) and architectures requiring parallelism far beyond current levels.

Further reading

Baede, **A.**, **D. Dent**, & **A. Hollingsworth**, 1976: The effect of arithmetic precision on some meteorological integrations. *ECMWF Technical Report No. 2.*

Düben, **P.D.**, **H. McNamara**, & **T.N. Palmer**, 2014: The use of imprecise processing to improve accuracy in weather & climate prediction. *J. Comput. Phys.*, **271**, 2–18.

Düben, **P.D.**, **F.P. Russell**, **X. Niu**, **W. Luk**, & **T.N. Palmer**, 2015: On the use of programmable hardware and reduced numerical precision in earth-system modeling. *J. Adv. Model. Earth Syst.*, **7**, 1393–1408.

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

The information within this publication is given in good faith and considered to be true, but ECMWF accepts no liability for error or omission or for loss or damage arising from its use.

doi:10.21957/avq2i3tp