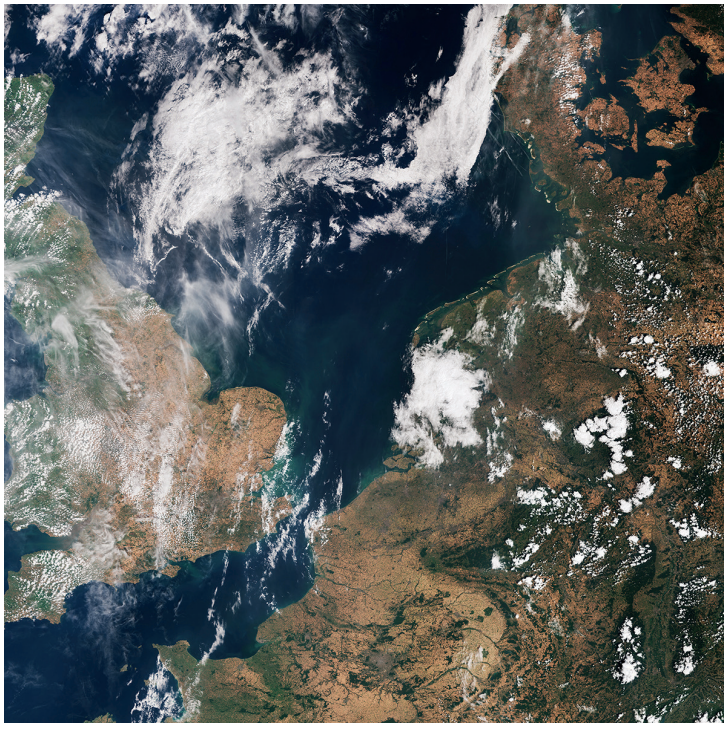
ECMWF Feature article

from Newsletter Number 157 – Autumn 2018

METEOROLOGY

Progress in using single precision in the IFS



Satellite image of Europe – 25 July 2018 © ESA

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

doi: 10.21957/ps2y9gfa2d

This article appeared in the Meteorology section of ECMWF Newsletter No. 157 – Autumn 2018, pp. 26-31.

Progress in using single precision in the IFS

Peter Düben, Michail Diamantakis, Simon Lang, Sami Saarinen, Irina Sandu, Nils Wedi, Tomas Wilhelmsson

Research carried out at the University of Oxford, Météo-France and ECMWF has shown that it is possible to significantly reduce the arithmetic precision of many of the calculations performed in numerical weather prediction models without compromising the quality of weather forecasts. 'Single precision' forecasts have the advantage of being computationally less expensive than traditional 'double precision' forecasts. Such efficiency savings will greatly facilitate the introduction of higher-resolution ensemble forecasts and other model improvements in line with ECMWF's Strategy to 2025.

A lot of work has gone into enabling the use of single precision in ECMWF's Integrated Forecasting System (IFS), with the result that the quality of single precision forecasts at the operational resolution is now comparable to that of double precision forecasts. The main remaining difference between single and double precision forecasts is a larger mass conservation error when single precision is used. The reasons for this have been found to be complex, but the error can be mitigated by using a global mass fixer. At ECMWF, single precision simulations have already proved useful to reduce computational cost significantly in research experiments. Work to prepare for the use of single precision operationally in the IFS is under way.

What is single precision?

When a weather forecast model runs on a supercomputer, physical parameters are represented as strings of bits that can be either 0 or 1. The precision at which a number can be represented depends on the number of bits that is used per variable. In the IFS, the default number of bits per real number has been 64 for the last few decades. This level of precision is called 'double precision'. It makes it possible to represent real numbers to a precision of at least 15 significant decimal digits. Numbers as large as 10308 and as small as 10-308 can be represented. In single precision, the number of bits to represent real numbers is reduced to 32. Precision for real numbers is reduced to seven significant decimal digits, with a number range between 10-38 and 1038. In general, the use of single precision instead of double precision speeds up simulations since less work needs to be done by the supercomputer. For uncoupled IFS simulations, this leads to a reduction in computing time of approximately 40%.

In the future, we aim to run standard forecasts with the IFS in single precision to improve computational efficiency while in principle keeping double precision for 4D-Var data assimilation. This includes forecasts for research purposes but also operational weather forecasts produced on ECMWF's next high-performance computing facility. See Box A for details on how single precision has been implemented in the IFS.

Why is single precision faster?

There are four reasons why single precision simulations are faster compared to double precision simulations:

1. Since data volume is reduced, more data can be stored closer to the processing unit (in memory and cache), and less waiting time and costly data transport is required.

2. The processing unit can perform more operations, with a speed-up by a factor of up to two. However, the size of any benefit depends on the extent to which the code is vectorised. Vectorisation is a style of computer programming in which operations are applied simultaneously to whole arrays instead of individual elements, with the number of parallel operations increased by a factor of two for single precision. Whether a significant ratio of the code is vectorised depends heavily on the compiler used.

3. Future supercomputers will use more and more processing units in parallel for a single simulation. The amount of information that needs to be shared between processors represents one of the most important bottlenecks for simulations. If single precision is used instead of double precision, the data volume that needs to be communicated between processors and compute nodes is halved.

4. For very large simulations, load balancing between compute nodes can be improved if overall data volume is reduced thanks to the use of single precision.

 doi: 10.21957/ps2y9gfa2d

Implementation of single precision in the IFS

A

In Fortran, KIND values define the number of bytes used to represent real numbers: 4 bytes = 32 bits correspond to single precision, 8 bytes = 64 bits correspond to double precision. KIND values are specified when real numbers are initialised at the beginning of programs, sub-routines or modules. In the IFS, precision is adjusted using a few global integer variables that define the KIND values for large groups of real numbers. There are three integer variables that are used to define the precision of most real number variables in the IFS:

- JPRB: This is the working precision that can be either double or single precision depending on the settings selected by the user.
- JPRM: These variables are always initialised in single precision.
- JPRD: These variables are always initialised in double precision.

JPRB is used for the overwhelming number of real numbers throughout the IFS model code. To change a simulation from double to single precision is, in principle, as easy as switching JPRB from 8 to 4 and changing some compiler options to define the use of single precision as default precision for variables that are initialised with no explicit specification of the KIND value. Starting from IFS Cycle 45r2 (a non-operational, technical cycle), it has been possible to choose the numerical precision and to start single precision simulations straight from prepIFS (using the 'Numerical precision' tab, under which users can pick a default precision and a precision for the 'FC' standard forecast job) with no need for any changes of the IFS branch. Jobs in single precision will automatically switch on the mass fixer.

If a local area in the code shows problematic behaviour if single precision is used, local variables can easily be upgraded from JPRB to JPRD to restore double precision locally. However, things become more complicated if relevant parameters are shared between subroutines, since this requires that the precision of information that is sent fits the precision of information that is received.

To make a special rule for single or double precision within the IFS code, an IF statement can be used to check whether JPRB is equal to JPRD. JPRB is equal to JPRD for double precision simulations and different for single precision simulations. This is, for example, useful if subroutines from libraries (such as LAPACK or BLAS) that are precision dependent are linked. The use of single precision to read GRIB input files or to write to GRIB output files, as well as MPI communication, is handled via interface blocks that pick the correct precision level automatically. The use of single precision will not change the precision of the GRIB data that is used for model I/O and data storage.

The IFS is rather complex with many different components that can have a very different computational footprint. This makes it difficult to carry out a reliable performance analysis. It is therefore not possible to make general statements on which of the four reasons listed above are the most important. Speed-ups depend on the hardware used; the model resolution; the MPI/OpenMP configuration and the number of processing units; the blocking of the code (using 'nproma'); the compiler; and other factors. It is possible for some model components to speed up by more than a factor of two if single precision is used (for example if expensive data operations suddenly fit into the limited size of the cache which stores data very close to the processing unit). For other model components, speed may hardly change (for example if performance is limited by the time certain information needs to travel from processor A to processor B, rather than the data volume). In general, for the IFS we have found a reduction in model-run time of approximately 40% for uncoupled simulations if single precision is used (*Váňa et al.*, 2016 and 2017). This result is consistent with results from other models (see for example *Nakano*, 2018).

Making single precision work in the IFS

The idea to use single precision in the IFS emerged from a research project at the University of Oxford carried out by Peter Düben and ECMWF Fellow Tim Palmer. Single precision was tested in the OpenIFS model, a portable version of the IFS that can be used for research projects at universities. It was shown that single precision simulations are possible and that the results are reasonable since differences between single and double precision simulations were smaller than the spread between ensemble members in ensemble simulations (see *Düben & Palmer*, 2014).

As a next step, single precision was introduced as an option into the IFS. A more extensive model evaluation revealed that single precision produced comparable results to double precision for ensemble simulations at about 50 km horizontal resolution (*Váňa et al.*, 2017). This work was carried out in close collaboration between ECMWF and Météo-France, which also successfully tested the use of single precision in global simulations.

doi: 10.21957/ps2y9gfa2d

Since then, a number of improvements to the single precision configuration at ECMWF have been made, and differences between the single precision configuration and the operational double precision configuration have been removed. This includes fixes in the Legendre transforms, the IO server, the coupling to the wave model WAM and the ocean model NEMO, the new radiation scheme, post-processing, the lake scheme, the vertical integration scheme of the dynamical core, the trigonometric grid information for simulations at very high resolution, and the fixing of several bugs that were not related to single precision but were identified thanks to the use of a different data layout in single precision. Many researchers at ECMWF have been involved in this process. As a result, today no additional code changes are required to run single precision forecast experiments in the IFS.

It should be noted that obtaining comparable results for single and double precision requires the use of double precision for some model components, including the pre-computation of Legendre polynomials and the operators of the vertical integration scheme before the time step loop is started. See Box B for details on which parts of the model code are sensitive to the use of single precision.

When do I need to be careful when using single precision? B

There are a couple of common operations that can cause problems with numerical precision. They should therefore be avoided whenever possible or fixed using locally enforced double precision (via JPRD, see Box A).

- If large numbers are multiplied or if a number is divided by a very small number, results may become larger than the largest number that can be represented at a certain level of precision (for example $>10^{38}$ for single precision). This will cause a number overflow and a crash of the model run. Rearranging the order of operations is often sufficient to avoid the multiplication of large numbers (for example $X^4/Y^6 \longrightarrow (X/Y)^4/Y^2$). If a divisor is very small and if there is a risk that the divisor may actually become zero, a number overflow can be avoided by adding a very small value (an epsilon) to the divisor.
- If numbers that are very similar in magnitude are subtracted from each other, several digits of precision can be lost in a single operation.
- It is possible for very small numbers that are added to large numbers to be rounded to zero. Even if the contribution of each summand on its own is not essential within a large sum, errors may be introduced when many small contributions are rounded to zero. It is therefore useful to begin sums over many numbers with the smallest number and to increase the size of the numbers that are added (a sum over pressure at model levels should, for example, start at the top of the model).
- If a very large or a very small number for which the exact value is not essential is required in the model code, it is important to use the intrinsic functions huge(1.0_JPRB) and epsilon(1.0_JPRB) to generate those numbers. This will adjust the value of numbers to the precision level that is used. Hard-coded numbers such as 10100 or 10-100 can cause a number overflow or will be rounded to zero if single precision is used.

doi: 10.21957/ps2y9gfa2d

Quality of single precision forecasts

In general, if single precision is implemented as described above, forecasts produced using single and double precision are very similar. The main difference between a single and a double precision simulation is a larger error in the conservation of total air mass for single precision simulations. While we have been able to reduce the magnitude of the error, the mass conservation error in single precision simulations still has an effect on forecast scores. We have identified three sources of the mass conservation error: the Legendre transformation, the vertical integration scheme and the semi-Lagrangian part of the model. The error in mass is fluctuating and can be both positive and negative with a small mean bias. Investigations have shown that reducing the error in mass conservation is not straightforward as it depends on a number of factors: it varies with resolution, is different for different initial conditions and even shows an annual cycle. However, the impact of the error in mass conservation can be mitigated by using a global mass fixer, which is cheap and easy to apply in a spectral model such as the IFS. If the mass fixer is switched on, differences in root-mean-square error between single precision and double precision geopotential height forecasts at the highest operational resolution (HRES) are mostly insignificant (Figure 1). However, even with the mass fixer switched on, there is currently still a degradation in some ensemble scores if single precision is used instead of double precision (Figure 2). Further investigations aimed at removing the remaining degradations are under way.

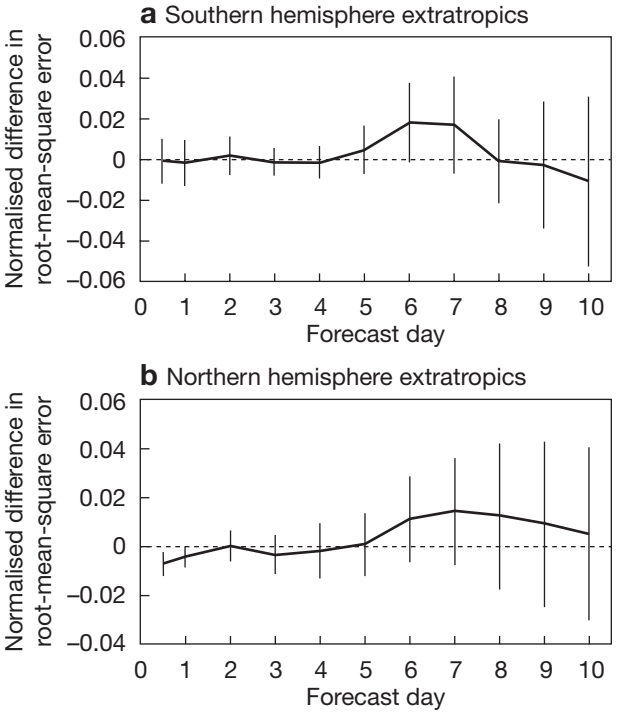
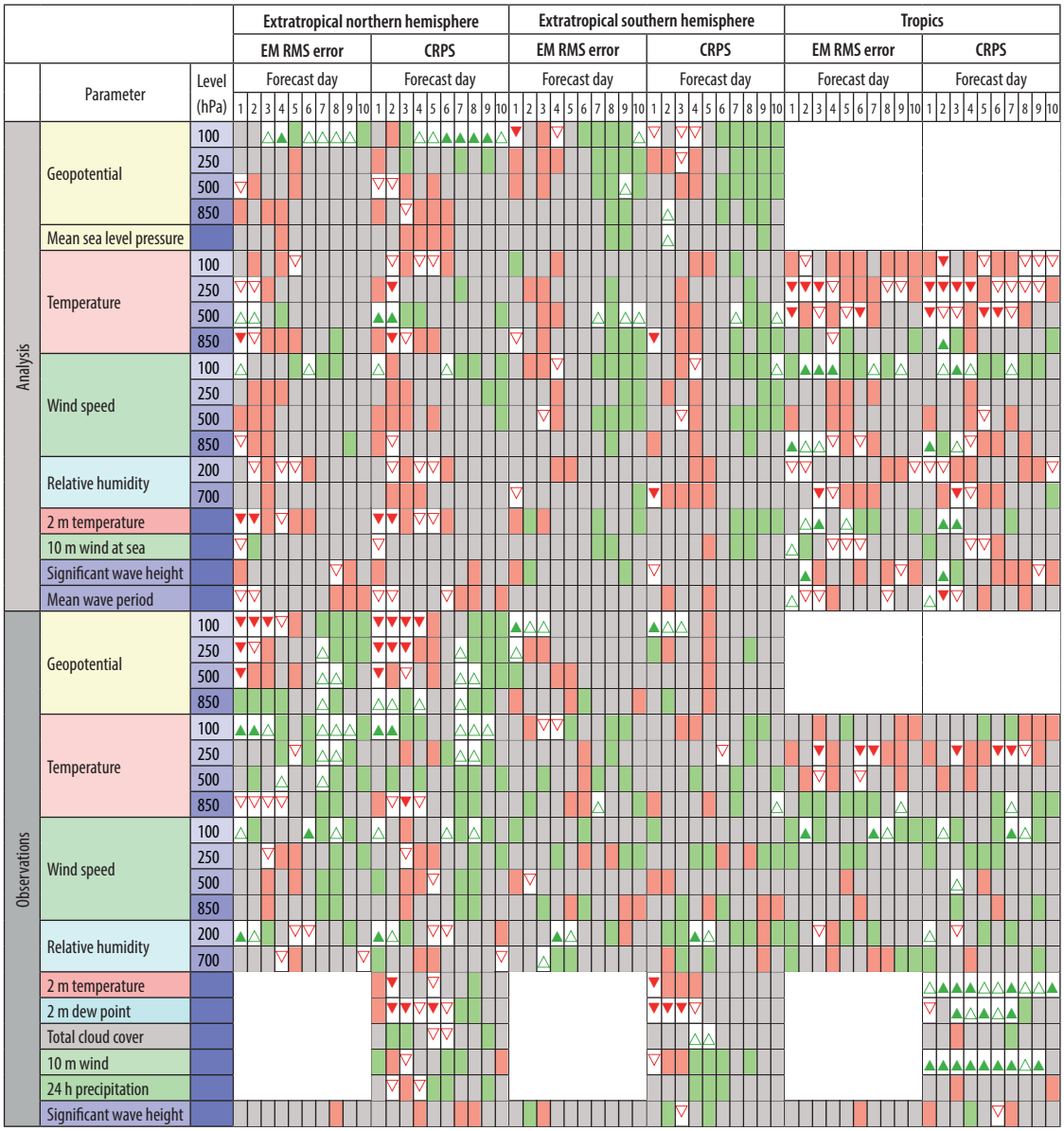


Figure 1 Normalised difference in the root-mean-square error for geopotential height at 500 hPa for a set of simulations in double and single precision at 9 km horizontal resolution (TCo1279) with 137 vertical levels for (a) the southern hemisphere extratropics (20°S to 90°S) and (b) the northern hemisphere extratropics (20°N to 90°N). For single precision simulations, the mass fixer was switched on. The difference was calculated as 'double precision' minus 'single precision' so that positive values indicate better results for single precision. The figures are based on the average of 45 simulations during January and February 2018. Vertical bars indicate the 95% confidence range.

doi: 10.21957/ps2y9gfa2d



Symbol legend: for a given forecast step...

- ▲ Single precision better than double precision statistically significant with 99.7% confidence
- △ Single precision better than double precision statistically significant with 95% confidence
- Single precision better than double precision statistically significant with 68% confidence
- No significant difference between single precision and double precision
- Single precision worse than double precision statistically significant with 68% confidence
- ▽ Single precision worse than double precision statistically significant with 95% confidence
- ▼ Single precision worse than double precision statistically significant with 99.7% confidence

Figure 2 Ensemble score card comparing single precision to double precision (both with mass fixer). Results are based on 45 ensemble simulations up to forecast day 10 with 10 ensemble members during June, July and August 2017 at 18 km (TCo639) resolution with 137 vertical levels.

Single precision research experiments

Before making use of single precision in research experiments, it is important to verify that single precision simulations respond to changes in model configuration in the same way as double precision simulations. To test this, we have performed a set of simulations during winter that use double precision or single precision in the standard IFS model configuration on the one hand and in a model configuration in which the orographic gravity wave and low-level blocking parametrization was switched off on the other. This parametrization accounts for interactions of mountains with the flow that cannot be resolved explicitly on a given grid.

doi: 10.21957/ps2y9gfa2d

Figure 3 shows the differences between the simulations. Differences between single precision and double precision with the parametrization switched on are very small, while differences between the simulations with and without orographic gravity wave and blocking are virtually identical in single and double precision.

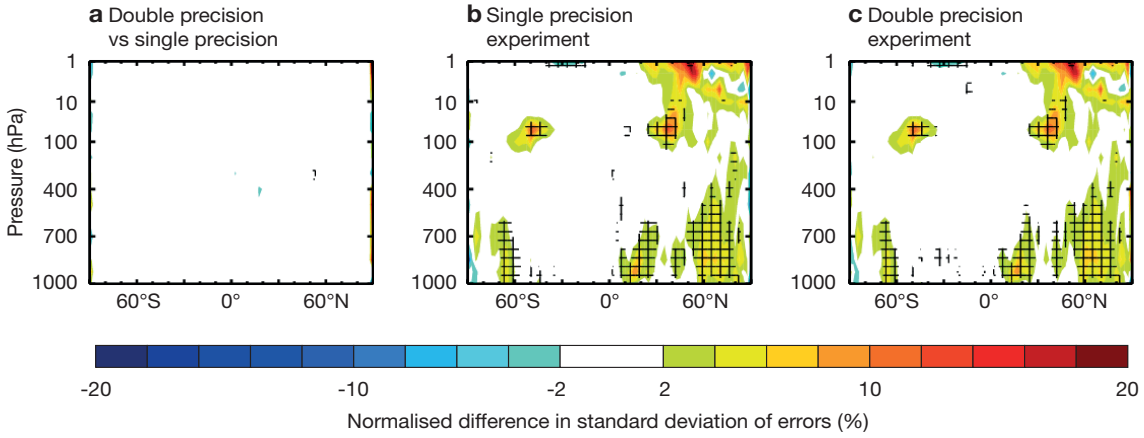


Figure 3 Normalised difference in standard deviation of errors against own analysis for vector components of horizontal winds for simulations at 25 km resolution (TCo399) and with 137 vertical levels after 24 hours, comparing (a) simulations with double precision with single precision simulations with mass fixer, (b) single precision simulations with and without gravity wave and low-level blocking parametrization and (c) double precision simulations with and without gravity wave and low-level blocking parametrization. The results are based on 30 forecasts with different starting times in December 2017. Cross-hatching indicates differences significant at the 95% level.

ECMWF aims to run ensembles at 5 km resolution by 2025. As the resolution increases beyond 10 km, some processes, such as convection or orographic drag, become resolved and the handover between parametrized and resolved processes poses a number of challenges. In order to prepare for future resolution upgrades and to make the most of the next supercomputer, it is therefore necessary to start testing the IFS at higher resolutions. However, testing the performance of the IFS at horizontal resolutions higher than that of the highest-resolution operational forecasts (9 km) is a very expensive exercise: the amount of information that needs to be available is very large and a very large amount of memory is needed to store the model state. Since the amount of memory that is available per compute node is limited, a large number of nodes is required for a single simulation. The overall performance of simulations is reduced since more data needs to be shared between processors and since the model needs to scale efficiently to a large number of processors.

Single precision will effectively reduce memory requirements by a factor of two, and this will have a very beneficial impact on the performance of simulations at very high resolution. We are therefore using single precision for tests in which the horizontal resolution is increased beyond the resolution of the deterministic operational forecasts. A series of two-day forecasts at horizontal resolutions ranging from 9 km to 1.25 km was performed for a day in August 2016 in the framework of the ESiWACE EU Horizon 2020 project, with both the IFS and the ICON model used by the German national meteorological service (DWD). The aim of these runs was to investigate scalability aspects in both models, but they are also very useful for understanding the challenges related to the representation of processes such as clouds, convection and precipitation in the ‘grey zone’. For example, Figure 4 shows that, at 9 and 5 km with ECMWF’s deep convection parametrization, the band of tropical rain over the Atlantic is too wide, but it has similar magnitude to the observed precipitation. In the 5, 2.5 and 1.25 km runs without the deep convection parametrization, the precipitation features become more realistic (the tropical band is narrower) but the rain is too intense. This suggests that work is needed both on the convection parametrization and its coupling with the dynamics in order to make the most of future resolution upgrades.

doi: 10.21957/ps2y9gfa2d

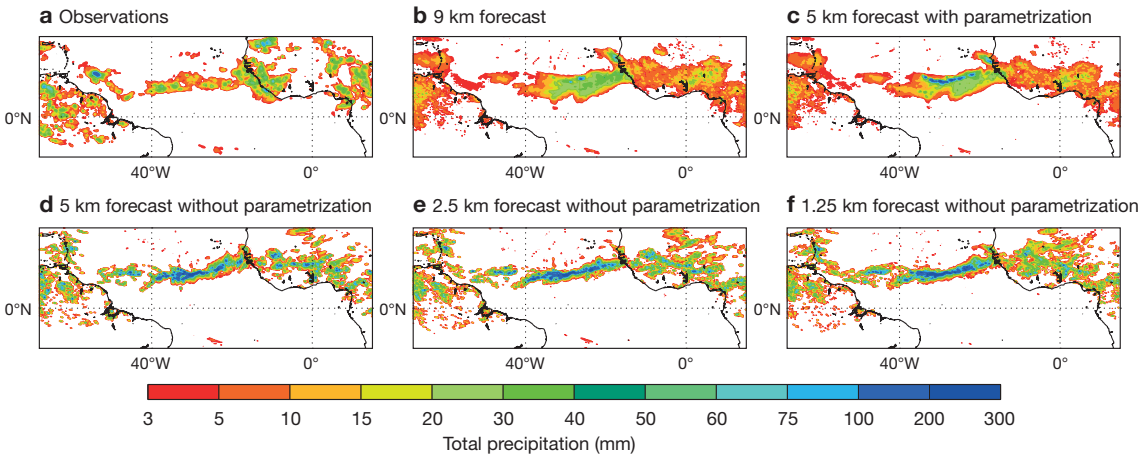


Figure 4 Total precipitation for 12 August 2016 over the tropical Atlantic according to (a) satellite observations (NASA's TRMM-3B42RT product), (b) a two-day 9 km horizontal resolution forecast (TCo1279), (c) a two-day 5 km horizontal resolution forecast (TCo1999) with deep convection parametrization, (d) a two-day 5 km horizontal resolution forecast without deep convection parametrization, (e) a two-day 2.5 km horizontal resolution forecast (TCo3999) without deep convection parametrization and (f) a two-day 1.25 km horizontal resolution forecast (TCo7999) without deep convection parametrization. The forecasts at TCo1999 and above were produced using single precision.

The future of single precision at ECMWF

The submission of ensemble simulations and analysis experiments that use the single precision version of the IFS will be possible straight from prepIFS from IFS Cycle 46r1, without any need for additional changes to the model configuration. However, the minimisation part of the 4D-Var data assimilation in the IFS and the tangent linear and adjoint models will remain in double precision since this part of the model has shown strong sensitivity to rounding errors in the past. To achieve optimal performance in standard research simulations using single precision, it may be useful to adjust the MPI/OpenMP configuration as well as code blocking (nproma).

In the near future, single precision will be tested for monthly and seasonal predictions as well as for atmospheric composition simulations. It is difficult to identify all differences between single and double precision in all aspects of the IFS. It will therefore be important for domain experts to have a more detailed look at the quality of single precision simulations in their specific area of expertise (such as land surface, ocean coupling, cloud physics, radiation, convection, stochastic parametrization schemes...) to identify any remaining differences.

In a second step, the use of single precision should also be tested in the NEMO ocean model. Preliminary tests on the use of single precision in NEMO at the Barcelona Supercomputing Centre are promising but more work is required.

In close collaboration with Tim Palmer's group at the University of Oxford, a reduction in numerical precision for weather forecasts beyond single precision is being investigated, such as the use of half precision (16 bit) arithmetic when calculating the Legendre transforms within the IFS, or a reduction in numerical precision when calculating dynamics at small spatial scales (*Hatfield et al.*, 2018; *Thornes et al.*, 2018).

The operational use of single precision will be a key element in moving towards the target of a 5 km ensemble set by ECMWF's Strategy to 2025. It will free up vital computational resources for forecast production and will thus maximise the benefits from the investment in ECMWF's next high-performance computing facility in Bologna from 2021.

doi: 10.21957/ps2y9gfa2d

Further reading

Düben, **P.D.** & **T.N. Palmer**, 2014: Benchmark Tests for Numerical Weather Forecasts on Inexact Hardware. *Mon. Wea. Rev.*, **142**, 3809–3829.

Hatfield, **S.**, **A. Subramanian**, **T. Palmer** & **P. Düben**, 2018: Improving Weather Forecast Skill through Reduced-Precision Data Assimilation. *Mon. Wea. Rev.*, **146**, 49–62.

Nakano, **M.**, **H. Yashiro**, **C. Kodama** & **H. Tomita**, 2018: Single Precision in the Dynamical Core of a Nonhydrostatic Global Atmospheric Model: Evaluation Using a Baroclinic Wave Test Case. *Mon. Wea. Rev.*, **146**, 409–416.

Thornes, **T.**, **P. Düben** & **T. Palmer**,, 2018: A Power Law for Reduced Precision at Small Spatial Scales: Experiments with an SQG Model. *Q J R Meteorol Soc*, doi: 10.1002/qj.3303.

Váňa, **F.**, **G. Carver**, **S. Lang**, **M. Leutbecher**, **D. Salmond**, **P. Düben** & **T. Palmer**, 2016: Single-precision IFS, *ECMWF Newsletter* **No. 148**, 20–23.

Váňa, **F.**, **P. Düben**, **S. Lang**, **T. Palmer**, **M. Leutbecher**, **D. Salmond** & **G. Carver**, 2017: Single Precision in Weather Forecasting Models: An Evaluation with the IFS. *Mon. Wea. Rev.*, **145**, 495–502.

doi: 10.21957/ps2y9gfa2d