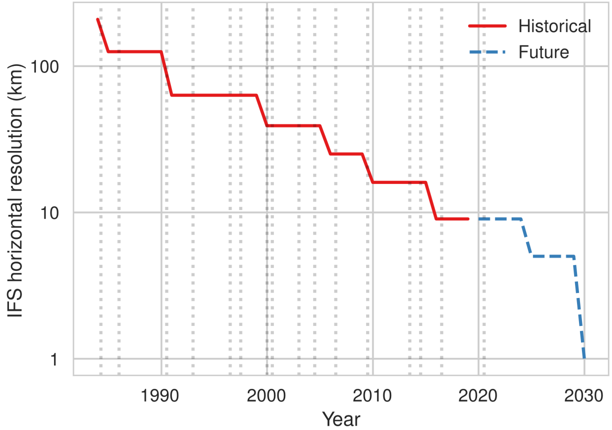
Mixed-Precision Arithmetic in Earth-System Modelling

Sam Hatfield, Peter Düben, Kristian Mogensen, Nils Wedi

samuel.hatfield@ecmwf.int

European Centre for Medium-Range Weather Forecasts

The quiet revolution



Tomorrow's workhorse supercomputers

Top500 #1 Fugaku, RIKEN R-CCS

Top500 #2 Summit, © ORNL and Carlos Jones

Outline

Part 1

- Floating-point numbers
- Why reduce precision?

Part 2

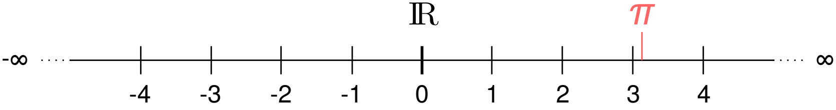
- Case studies
- Single-precision at ECMWF

Floating-point numbers

Real numbers on computers

Numerical models use **real number arithmetic**

We therefore need a way to map a number from the real number line:



to a finite bitstring:

$\pi \approx$

0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	1	0	0	0	0	1	1	1	1	1	1	0	1	1
0	1	0	1	0	1	0	0	0	1	0	0	0	1	0	0	0	0	1	0	1	1	0	1	0	0	0	1	1	0	0	0

64 bits

so that we can do that arithmetic on a computer

On a finite computer this is **inherently imperfect**

The obvious way: fixed-point numbers

We can create a crude real number format from an integer simply by placing a "binary point" somewhere, e.g.

```
 10110110 = 182    8 bit integer
10110.110 = 22.75  8 bit fixed-point number (binary point at 5th place)
```

- Advantages
  - We can reuse the integer arithmetic chip (**fast**)
- Disadvantages
  - **Very** low precision (depending on the position on the number line)
  - Limited range

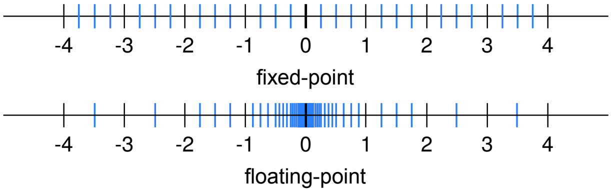
A better way: floating-point numbers

Instead we use **floating-point** numbers:

$$x = \underbrace{\text{fixed-point number}}_{\text{significand}} \times 2^{\overbrace{\text{integer}-\text{bias}}^{\text{exponent}}}$$

e.g. for a 64 bit “float”

$$\pi = 1.5707963267948966 \times 2^{1024-1023}$$



Floating-point standard

Three formats according to IEEE:

	sign	exponent	significand	
double	0	10000000000 (11 bits)	10010010000111111011 01010100010001000010110100011000 (52 bits)	= 3.141592653589793
single	0	10000000 (8 bits)	10010010000111111011011 (23 bits)	= 3.1415927
half	0	10000 (5 bits)	1001001000 (10 bits)	= 3.14

■ sign ■ exponent ■ significand

Floating-point properties

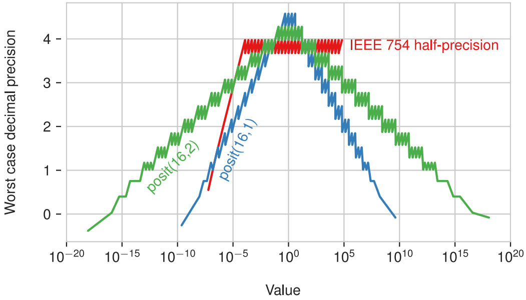
You should keep in mind:

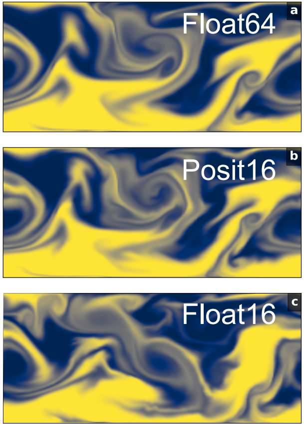
- **Machine epsilon**: smallest number that can be added to 1 to produce a different number (Fortran: `EPSILON`)
- **Smallest (non-subnormal) number**: smallest representable number in the "normal" range (Fortran: `TINY`)
- **Largest number**: largest representable number (Fortran: `HUGE`)

Precision	double	single	half
`EPSILON(x)`	$2.220446049250313 \times 10^{-16}$	1.1920929×10^{-7}	0.000977
`TINY(x)`	$2.2250738585072014 \times 10^{-308}$	$1.1754944 \times 10^{-38}$	0.00006104
`HUGE(x)`	$1.797693134862315 7\times 10^{308}$	3.4028235×10^{38}	65504 !

Digression: alternatives to floats

Could “posits” replace floats?

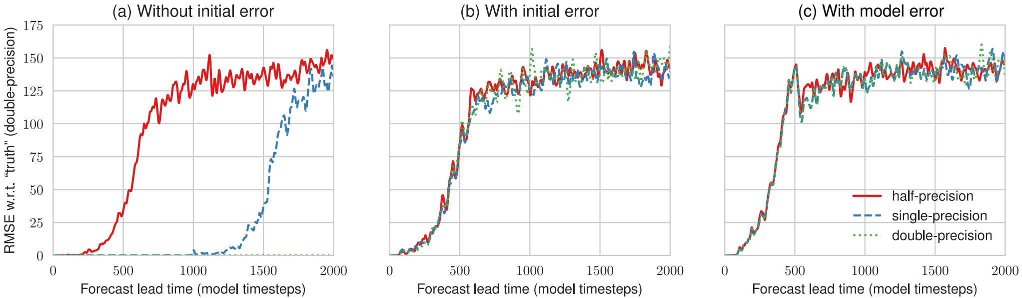




[Klöwer et al., 2020]

Why reduce precision?

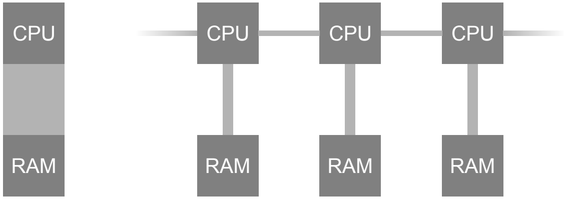
Initial and model error



Lorenz '63 example

A communication/memory-bound world

Basically all models are now **memory-bound**



What computers used to look like

What they look like now

Reducing precision effectively **increases the communication bandwidth**, thereby accelerating models

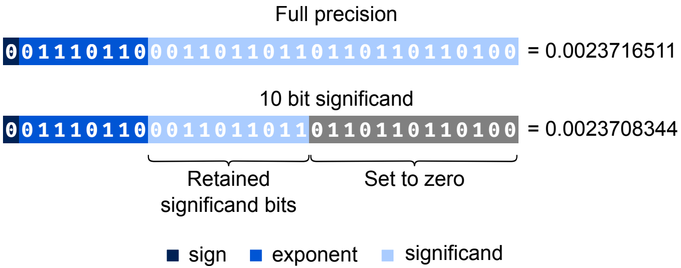
Case studies

Emulating reduced-precision computations

Most hardware only supports **double** and **single**-precision arithmetic

How do we assess feasibility of reducing precision without having to port to GPUs etc.?

Software emulation!

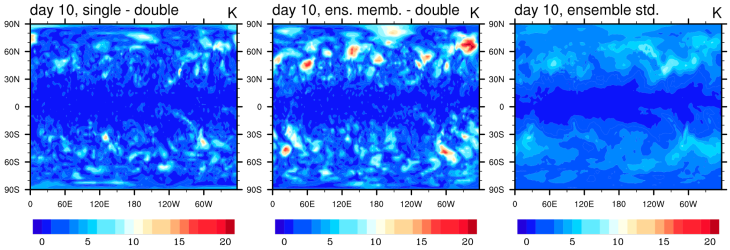


[Dawson and Düben, 2017]

Downside: we can only estimate computational cost savings

Single-precision in a realistic atmospheric model

Compare double-precision with single-precision in OpenIFS



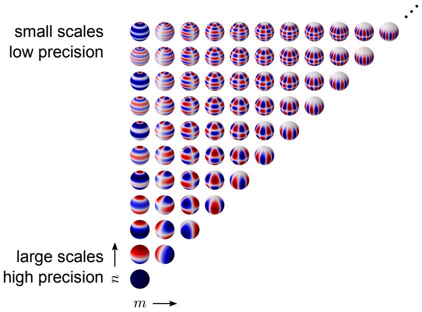
850 hPa temperature difference

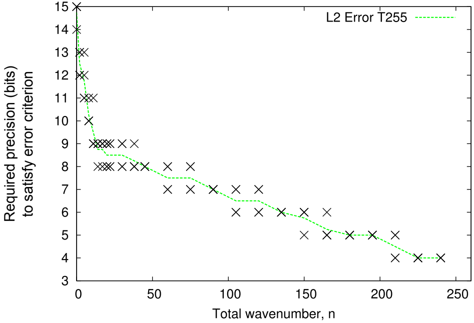
[Düben and Palmer, 2014]

diff(single, double) $\approx$ std(ensemble) $<$ diff(double member, double mean)

Scale-selective precision (1)

High-precision for large scales, low-precision for small scales?

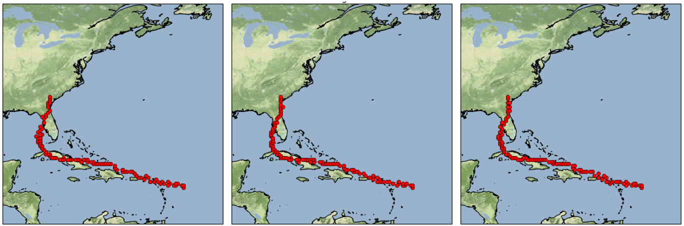




[Chantry et al., 2018]

Scale-selective precision (2)

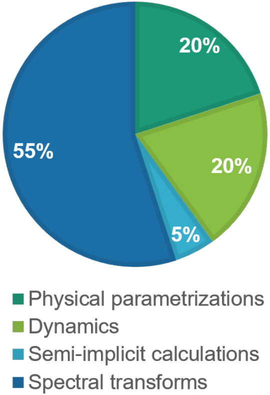
double single scale-selective



Hurricane Irma core position

Average precision of “scale-selective” (across all wavenumbers): **8.6 significand bits**

Reduced-precision Legendre transforms (1)

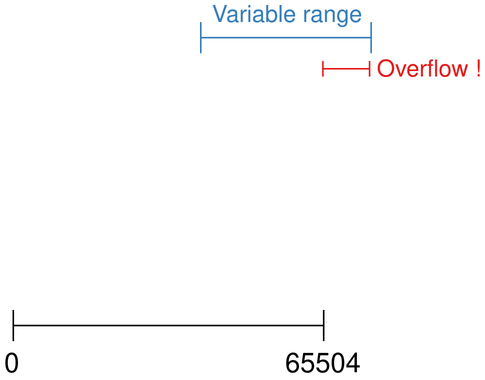


Profile of IFS at TCo7999 ($\sim$1.5 km)

- Target the spectral transforms to accelerate high-resolution IFS simulations
- GPUs allow **half-precision** or **Tensor Core** matrix multiplications

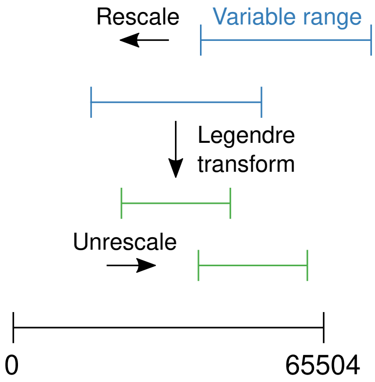
Reduced-precision Legendre transforms (2)

The largest half-precision number is **65504**
So we must **rescale** variables to avoid overflows



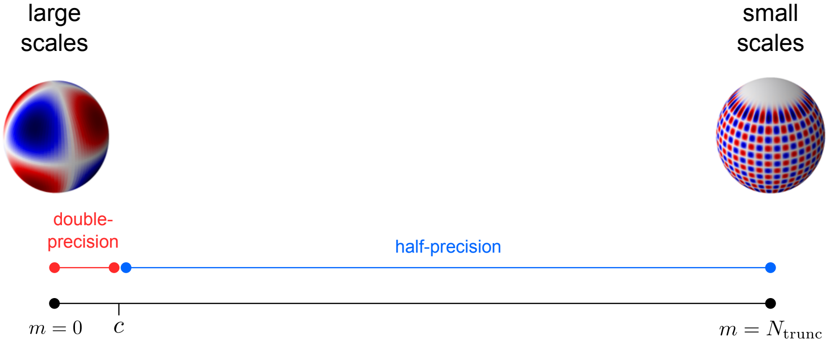
Reduced-precision Legendre transforms (2)

The largest half-precision number is **65504**
So we must **rescale** variables to avoid overflows



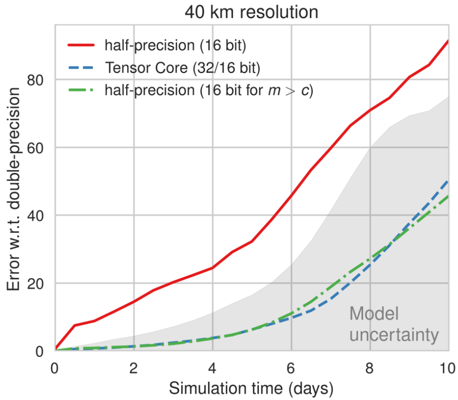
Reduced-precision Legendre transforms (3)

Half-precision rounding errors also cause problems for the **largest scale** calculations



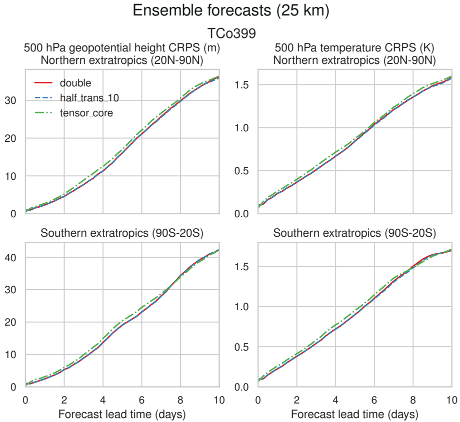
So keep them at double-precision

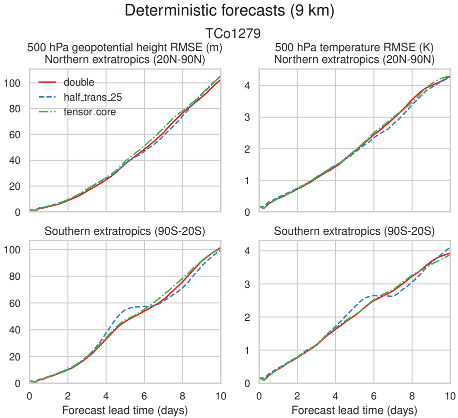
Reduced-precision Legendre transforms (4)



- "Model uncertainty" = double-precision with SPPT
- error(half-precision) $<$ model uncertainty **if** we protect the first c (in this case 10) modes

Reduced-precision Legendre transforms (5)





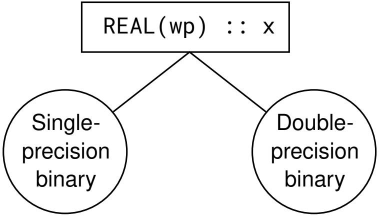
[Hatfield et al., 2019]

Single-precision at ECMWF

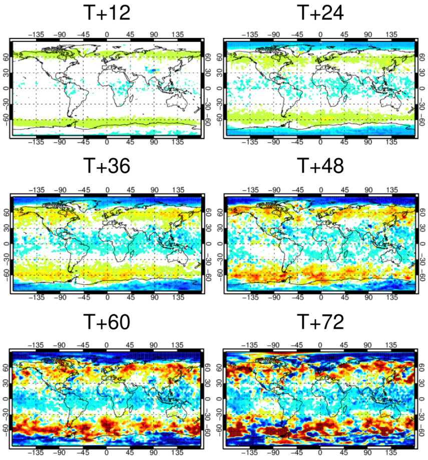
Single-precision ocean and atmosphere

ECMWF's philosophy on single-precision:

- Same source code for double and single-precision
- Upgrade specific parts to double-precision where necessary
- Roadmap: **fully single-precision coupled forecasts**

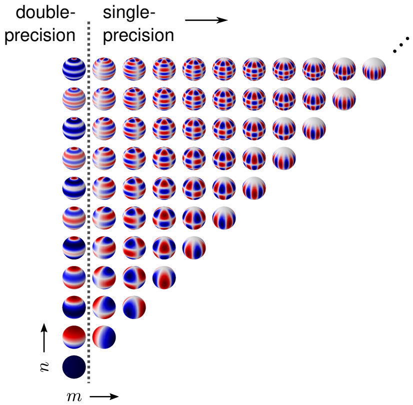


Single-precision in the atmosphere (development issues)

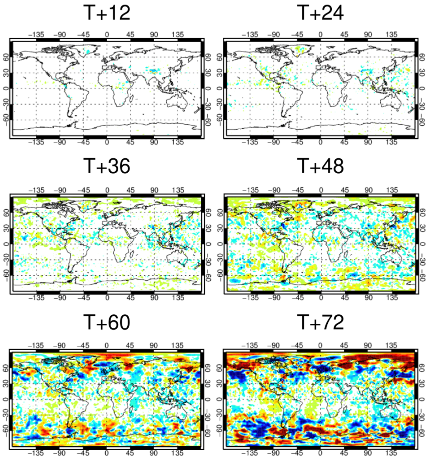


Change in mean sea-level-pressure error SP vs. DP

Promote zeroth mode to double-precision:

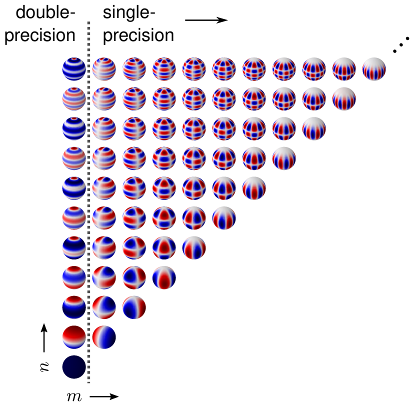


Single-precision in the atmosphere (development issues)



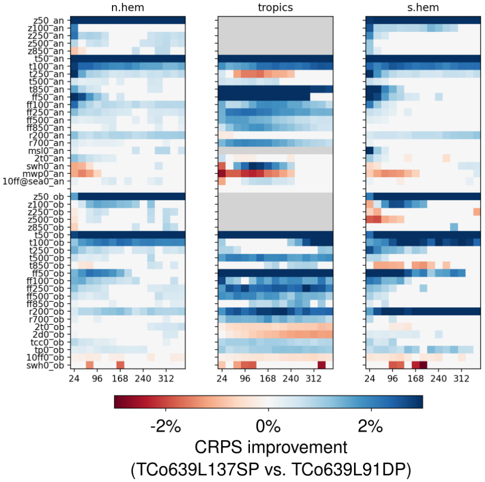
Change in mean sea-level-pressure error SP vs. DP

Promote zeroth mode to double-precision:



Single-precision in the atmosphere (status)

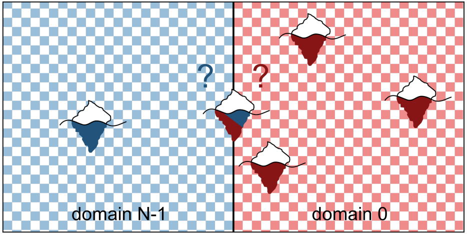
- Single-precision atmosphere **operational default** for HRES and ENS (not DA) from Cy47r2 onwards ($\sim$ January 2021)
- **1.7$\times$ times speed-up** compared with double-precision
- Allows free upgrade from 91 to 137 levels $\rightarrow$ improvement in forecast skill



Next: single-precision in the ocean

- NEMO now coupled to IFS in **all forecast products**
- 20% of total cost of EPS but **60% cost of seasonal system**
- Develop **single-precision** capability in NEMO, building on work at BSC [Tintó Prims et al., 2019]
- Single-precision in sea-ice and icebergs novel

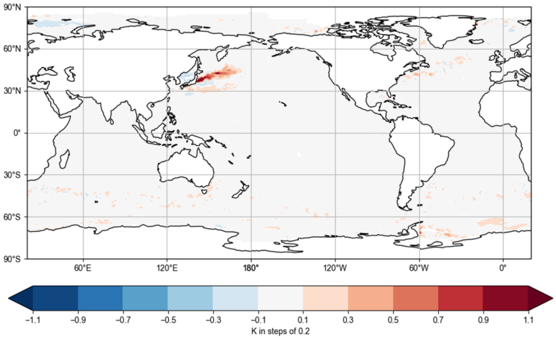
Example bug: orphaned icebergs



Affects mostly single-precision **but also double-precision**

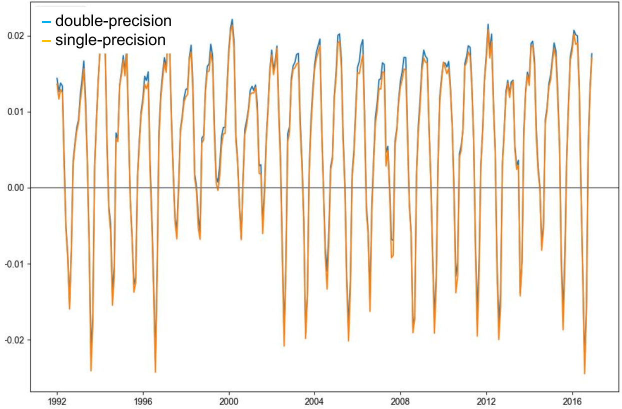
Single-precision in the ocean (status)

- Single-precision NEMO stable at 0.25° resolution with sea-ice and icebergs (operational configuration)
- Speed-up: **up to 1.7**× (depends on I/O)



Change in sea-surface temperature RMSE single vs. double-precision

Single-precision in the ocean (status)



Northern Hemisphere sea-ice concentration bias

Conclusion

- Supercomputing is undergoing a **paradigm shift**
- **Precision** will become an additional “knob” for **cost/accuracy** trade-off
- Earth-System modelling is an ideal application for exploring precision:
  - We have **model error** and **initial error**
  - Our models are **memory/communication-bound**
- Single-precision will be used operationally in the IFS **starting next year**
- Single-precision ocean and coupled modelling under testing

References

Chantry, M., Thornes, T., Palmer, T., and Düben, P. (2018).
Scale-Selective Precision for Weather and Climate Forecasting.
Monthly Weather Review, 147(2):645–655.

Dawson, A. and Düben, P. D. (2017).
Rpe v5: An emulator for reduced floating-point precision in large numerical simulations.
Geoscientific Model Development, 10(6):2221–2230.

Düben, P. D. and Palmer, T. N. (2014).
Benchmark Tests for Numerical Weather Forecasts on Inexact Hardware.
Monthly Weather Review, 142(10):3809–3829.

Hatfield, S., Chantry, M., Düben, P., and Palmer, T. (2019).
Accelerating High-Resolution Weather Models with Deep-Learning Hardware.
In *Proceedings of the Platform for Advanced Scientific Computing Conference - PASC '19*, pages 1–11. ACM Press.

Hatfield, S., Düben, P., Chantry, M., Kondo, K., Miyoshi, T., and Palmer, T. (2018).
Choosing the optimal numerical precision for data assimilation in the presence of model error.
Journal of Advances in Modeling Earth Systems.

Klöwer, M., Düben, P. D., and Palmer, T. N. (2020).
Weather and climate models in 16-bit arithmetics : Number formats , error mitigation and scope.
Journal of Advances in Modeling Earth Systems (submitted).

Tintó Prims, O., Acosta, M. C., Moore, A. M., Castrillo, M., Serradell, K., Cortés, A., and Doblas-Reyes, F. J. (2019).
How to use mixed precision in Ocean Models.
Geoscientific Model Development Discussions, pages 1–21.