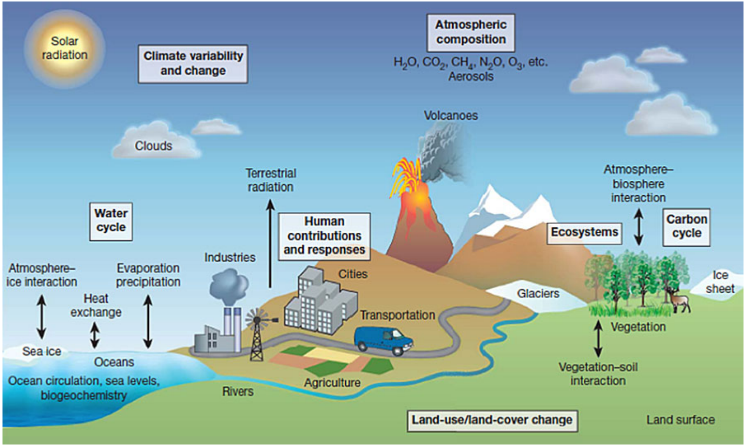
A detailed precision analysis for weather and climate models

Peter Düben

University Research Fellow of the Royal Society

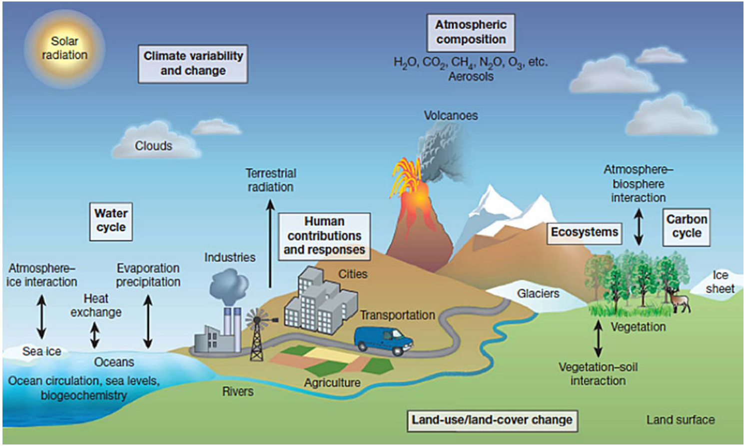
European Centre for Medium-Range Weather Forecasts (ECMWF)

Predicting weather and climate: Why is it so hard?



www.gfdl.noaa.gov

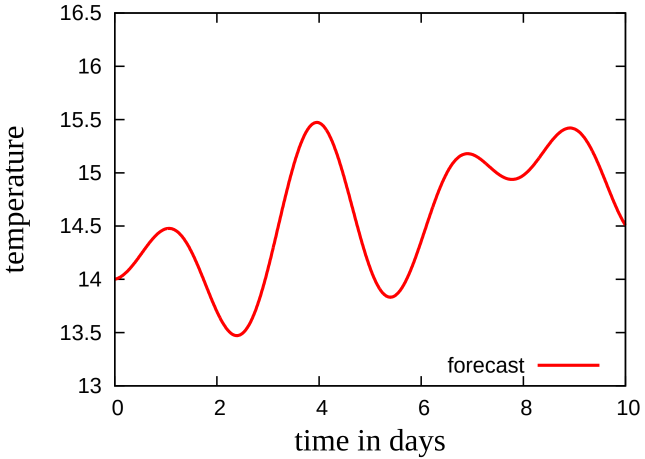
Predicting weather and climate: Why is it so hard?



www.gfdl.noaa.gov

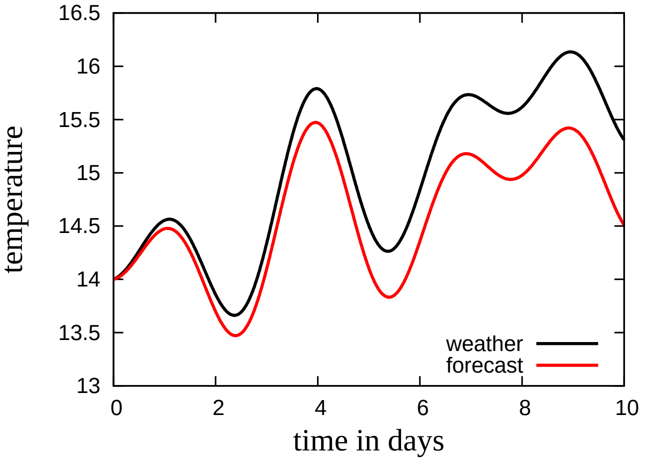
The Earth System is complex, huge and chaotic and we do not have sufficient resolution to resolve all important processes.

Ensemble forecasts to prediction model uncertainty



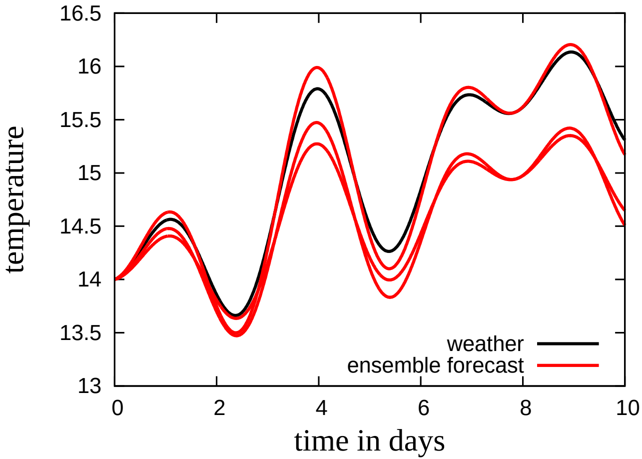
How do we know if we are wrong?

Ensemble forecasts to prediction model uncertainty



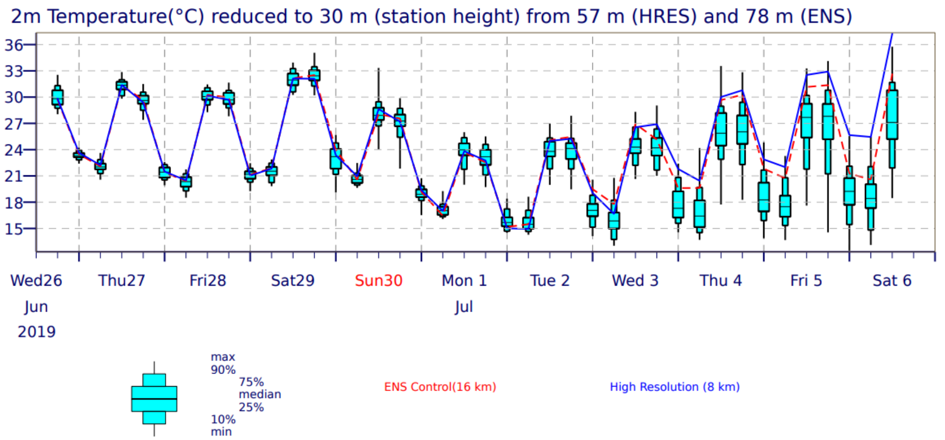
How do we know if we are wrong?

Ensemble forecasts to prediction model uncertainty

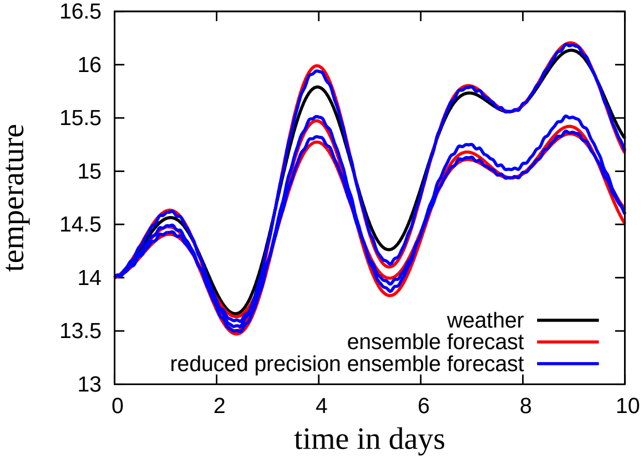


The ensemble spread holds information about forecast uncertainty.

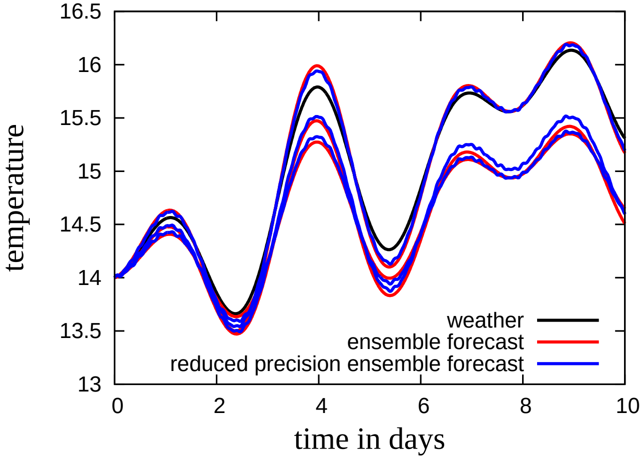
Ensemble forecasts to prediction model uncertainty



Accuracy adjusted to predictability



Accuracy adjusted to predictability



We can use the ensemble spread to adjust numerical precision.

Less numerical precision → more computing power

Double precision (64 bits) is used almost exclusively in weather and climate modelling.

Less numerical precision $\rightarrow$ more computing power

Double precision (64 bits) is used almost exclusively in weather and climate modelling.

Reduce numerical precision

$\rightarrow$ lower power, higher performance.

$\rightarrow$ higher resolution or increased complexity.

$\rightarrow$ more accurate predictions of future weather and climate.

Less numerical precision $\rightarrow$ more computing power

Double precision (64 bits) is used almost exclusively in weather and climate modelling.

Reduce numerical precision

$\rightarrow$ lower power, higher performance.

$\rightarrow$ higher resolution or increased complexity.

$\rightarrow$ more accurate predictions of future weather and climate.

Temperature in Paris:
double precision (64 bits): 14.561192512512207°C
single precision (32 bits): 14.5611925°C
half precision (16 bits): 14.5625°C

Less numerical precision $\rightarrow$ more computing power

Double precision (64 bits) is used almost exclusively in weather and climate modelling.

Reduce numerical precision

$\rightarrow$ lower power, higher performance.

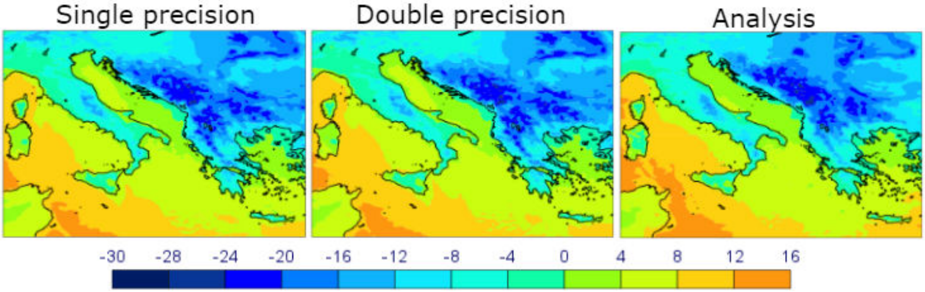
$\rightarrow$ higher resolution or increased complexity.

$\rightarrow$ more accurate predictions of future weather and climate.

Temperature in Paris:
double precision (64 bits): 14.561192512512207°C
single precision (32 bits): 14.5611925°C
half precision (16 bits): 14.5625°C

But can we really do it? And how far can we go?

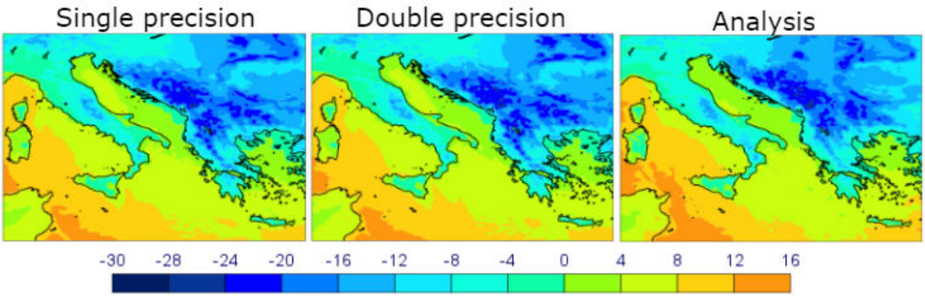
ECMWF's weather forecast model in single precision



- Forecast quality in double and single precision is almost identical.
- 40% reduction of run time.
- Benefit for global simulations at cloud-resolving resolution.

Düben and Palmer MWR 2014; Váňa, Düben et al. MWR 2017; Düben et al. ECMWF Newsletter 2018

ECMWF's weather forecast model in single precision



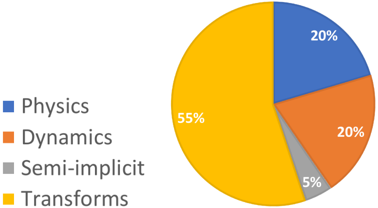
- Forecast quality in double and single precision is almost identical.
- 40% reduction of run time.
- Benefit for global simulations at cloud-resolving resolution.

Düben and Palmer MWR 2014; Váňa, Düben et al. MWR 2017; Düben et al. ECMWF Newsletter 2018

Can we go lower than single precision?

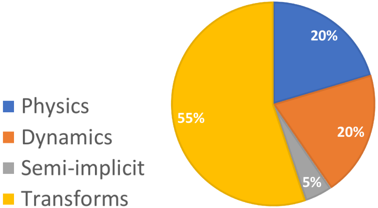
Machine learning hardware for fast simulations with low precision

Relative cost for model components for a non-hydrostatic model at 1.45 km resolution:



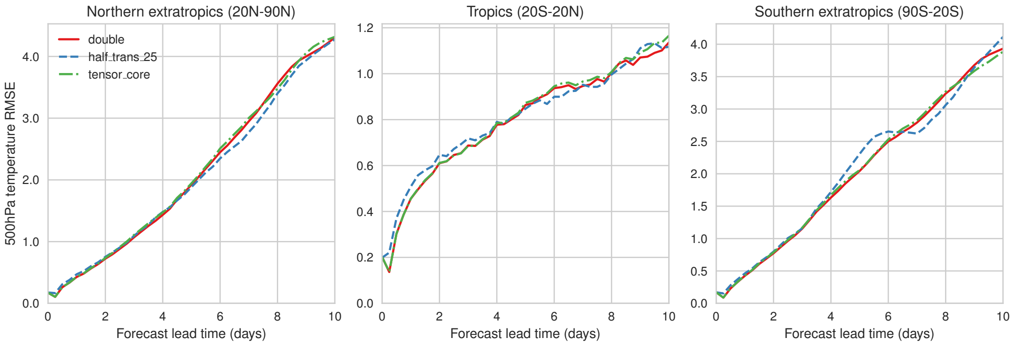
Machine learning hardware for fast simulations with low precision

Relative cost for model components for a non-hydrostatic model at 1.45 km resolution:



- The Legendre transform is the most expensive kernel. It consists of a large number of standard matrix-matrix multiplications.
- If we can re-scale the input and output fields, we can use half precision arithmetic.
- Tensor Cores on NVIDIA Volta GPUs are optimised for half-precision matrix-matrix calculations with single precision output. 7.8 TFlops for double precision vs. 125 TFlops for half precision on the Tensor Core.

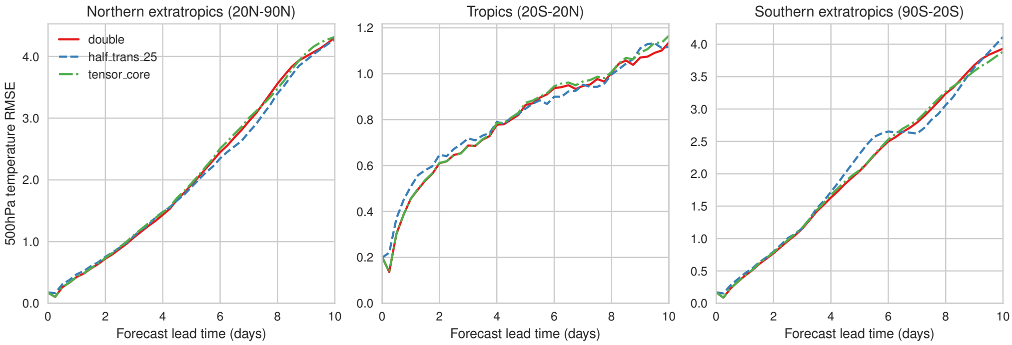
Half precision Legendre Transformations



Root-mean-square error for Z500 at 9 km resolution averaged over multiple start dates.

Hatfield, Chantry, Dueben, Palmer **Best Paper Award PASC2019**.

Half precision Legendre Transformations



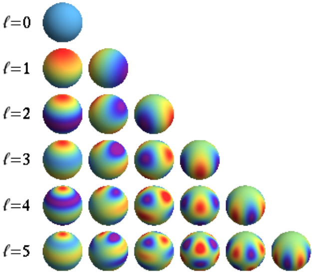
Root-mean-square error for Z500 at 9 km resolution averaged over multiple start dates.

Hatfield, Chantry, Dueben, Palmer **Best Paper Award PASC2019**.

The simulations are using an emulator to reduce precision.

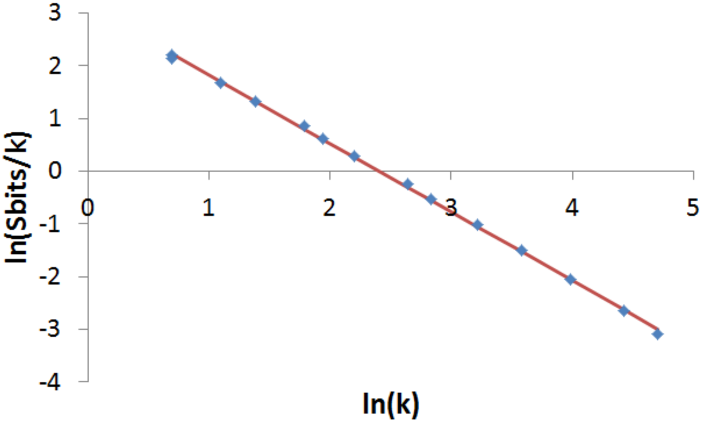
Dawson and Dueben GMD 2017

A scale-selective approach to adjust precision



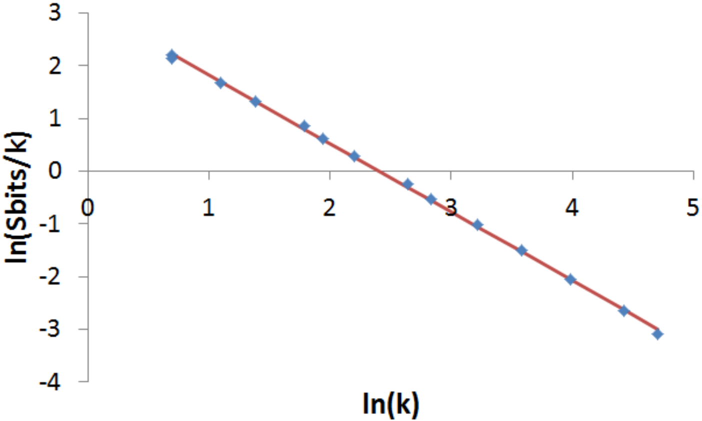
- Spectral models allow to treat different scales at different precision.
- We can reduce precision when calculating the small scales.
- This is intuitive due to the high inherent uncertainty in small scale dynamics (parametrisation, viscosity, data-assimilation,...).
- The smallest scales are most expensive.

A scale-selective approach



A scale-dependent reduction in precision for the surface quasi-geostrophic equations.

A scale-selective approach

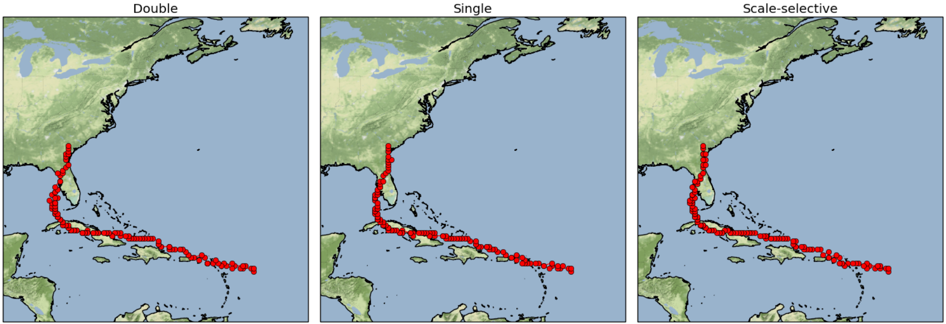


A scale-dependent reduction in precision for the surface quasi-geostrophic equations.

Forecast simulations confirm that a scale-selective approach is much more efficient than a uniform precision reduction.

Thornes, Düben and Palmer QJRMS 2017, Thornes, Düben and Palmer QJRMS 2018

A scale-selective approach: Track of Hurricane Irma



- Simulations with OpenIFS at 40 km resolution.
- The scale-selective simulation is using scale-selective precision in spectral space. An average of 8.6 bits is used for the significand.

Chantry, Thornes, Palmer, Dueben MWR 2019

Rounding errors adjusted to model error

We want rounding errors to be approximately equal to model errors.

Rounding errors adjusted to model error

We want rounding errors to be approximately equal to model errors.

The uncertainty of initial conditions provides the level for precision at the beginning of the forecast.

Rounding errors adjusted to model error

We want rounding errors to be approximately equal to model errors.

The uncertainty of initial conditions provides the level for precision at the beginning of the forecast.

Error growth of errors in initial conditions is roughly exponential.

Rounding errors adjusted to model error

We want rounding errors to be approximately equal to model errors.

The uncertainty of initial conditions provides the level for precision at the beginning of the forecast.

Error growth of errors in initial conditions is roughly exponential.

Rounding errors will decrease exponentially with the number of bits.

Rounding errors adjusted to model error

We want rounding errors to be approximately equal to model errors.

The uncertainty of initial conditions provides the level for precision at the beginning of the forecast.

Error growth of errors in initial conditions is roughly exponential.

Rounding errors will decrease exponentially with the number of bits.

$\rightarrow$ Precision should be reduced linearly with forecast lead time (proportional to the leading Lyapunov exponent).

Rounding errors adjusted to model error

We want rounding errors to be approximately equal to model errors.

The uncertainty of initial conditions provides the level for precision at the beginning of the forecast.

Error growth of errors in initial conditions is roughly exponential.

Rounding errors will decrease exponentially with the number of bits.

$\rightarrow$ Precision should be reduced linearly with forecast lead time (proportional to the leading Lyapunov exponent).

This would reduce cost by approximately a factor of two.

Rounding errors adjusted to model error

We want rounding errors to be approximately equal to model errors.

The uncertainty of initial conditions provides the level for precision at the beginning of the forecast.

Error growth of errors in initial conditions is roughly exponential.

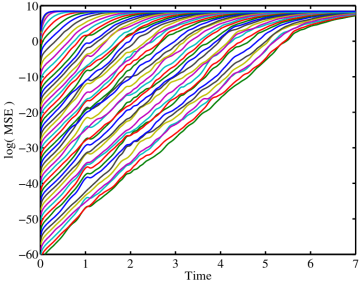
Rounding errors will decrease exponentially with the number of bits.

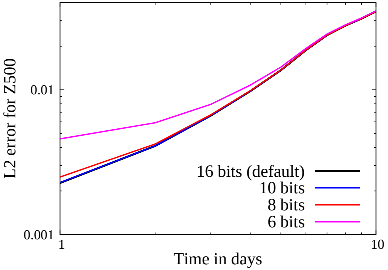
$\rightarrow$ Precision should be reduced linearly with forecast lead time (proportional to the leading Lyapunov exponent).

This would reduce cost by approximately a factor of two.

Limitations: Linear error growth of model error and seasonal predictions.

Rounding errors adjusted to model error

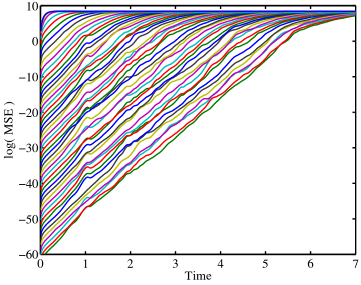


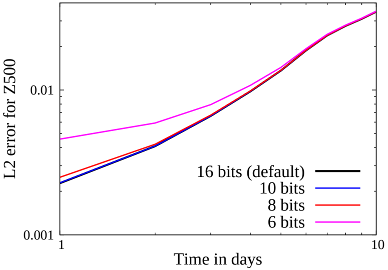


Left: Logarithm of the Mean Squared Error for simulations with Lorenz’95.

Right: Error for ECMWF data at different levels of precision.

Rounding errors adjusted to model error



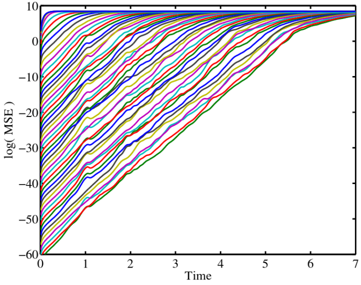


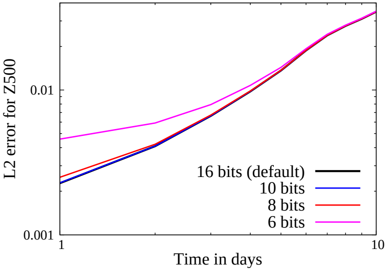
Left: Logarithm of the Mean Squared Error for simulations with Lorenz’95.

Right: Error for ECMWF data at different levels of precision.

Rounding errors can be linked to the level of model error.

Rounding errors adjusted to model error





Left: Logarithm of the Mean Squared Error for simulations with Lorenz'95.

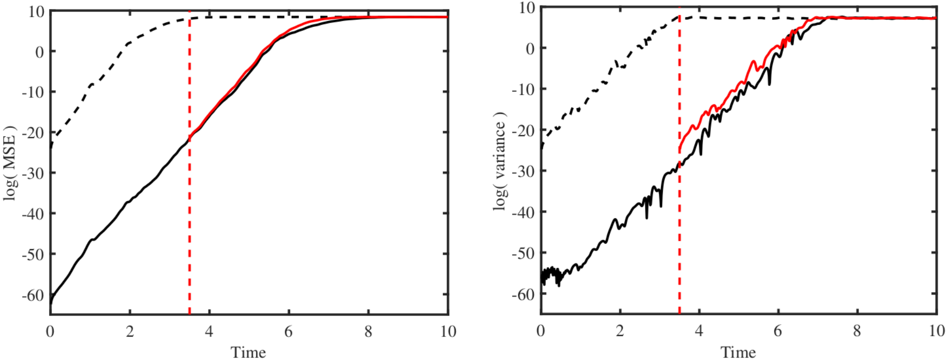
Right: Error for ECMWF data at different levels of precision.

Rounding errors can be linked to the level of model error.

Precision should be reduced with forecast lead time.

Düben et al. JAMES 2015, Cooper, Düben et al. submitted to MWR

To use verificarlo to diagnose precision reduction

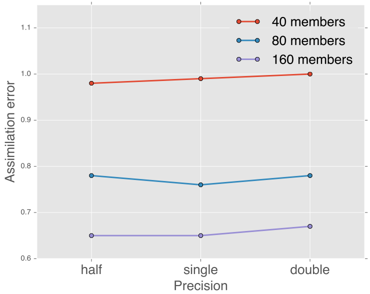


Left: Forecast error with reduced precision emulator (Dawson and Dueben GMD 2017).

Right: Error propagation diagnosed by Verificarlo.

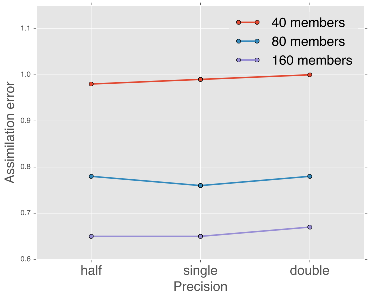
Cooper, Düben et al. submitted to MWR

Data assimilation with reduced precision



Data assimilation in Lorenz'95 using an Ensemble Kalman filter. Hatfield, Dueben, Palmer JAMES 2018

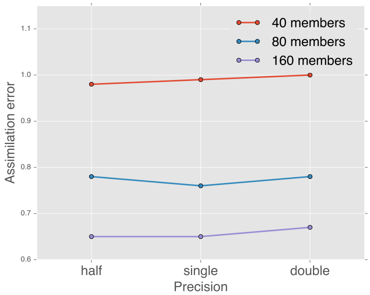
Data assimilation with reduced precision



Data assimilation in Lorenz'95 using an Ensemble Kalman filter. Hatfield, Dueben, Palmer JAMES 2018

A large ensemble at low precision is better than a small ensemble at high precision.

Data assimilation with reduced precision

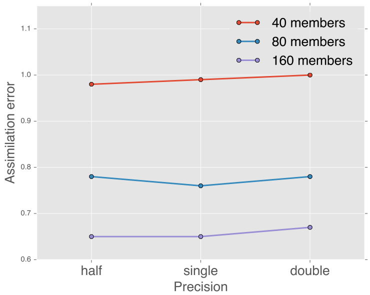


Data assimilation in Lorenz'95 using an Ensemble Kalman filter. Hatfield, Dueben, Palmer JAMES 2018

A large ensemble at low precision is better than a small ensemble at high precision.

We gain almost one "day" in terms of predictability.

Data assimilation with reduced precision



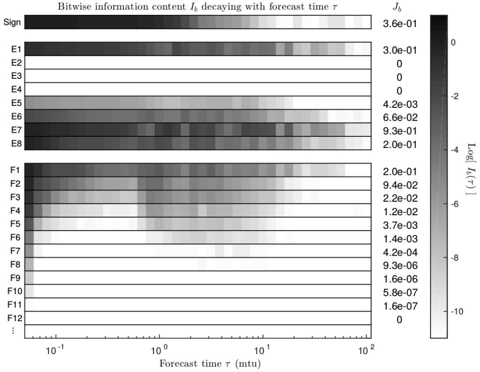
Data assimilation in Lorenz'95 using an Ensemble Kalman filter. Hatfield, Dueben, Palmer JAMES 2018

A large ensemble at low precision is better than a small ensemble at high precision.

We gain almost one "day" in terms of predictability.

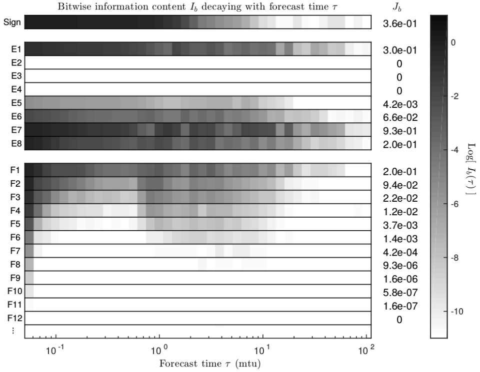
However, 4DVar data assimilation may be more difficult...

Bitwise information content and predictability



Information content of bits for a Lorenz'63 model using a single long term integration and Shannon information theory.

Bitwise information content and predictability



Information content of bits for a Lorenz'63 model using a single long term integration and Shannon information theory.

It is possible to identify information content of individual bits and their impact on predictability into the future.

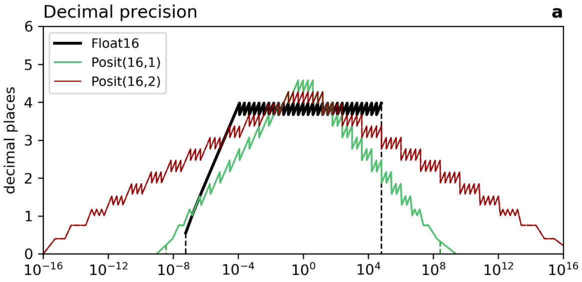
Jeffress, Düben and Palmer Proc. R. Soc. A 2017

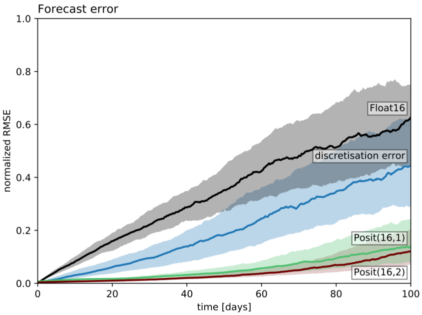
What number format to use?

16 bits is not much so you may need to show some flexibility and use Posits.

What number format to use?

16 bits is not much so you may need to show some flexibility and use Posits.





Left: Dynamic number ranges of 16 bit Posit formats and 16 bit half precision floats.
Right: Forecast error for a shallow water model if reduced precision is used.

Kloewer, Düben and Palmer CONGA 2019

Reduce precision in weather and climate models

What we still need:

- Tools that allow an automated search for the optimal precision level when non-linear feedbacks are present.
- A basic understanding how to formulate models to minimize numerical precision (re-scaling of equations, perturbation approaches, multi-grid solvers...).
- Tools to predict a performance increase from a precision reduction for a given hardware.
- Information how future hardware and hardware co-design will look like (CPUs, GPUs, TPUs, FPGAs, ASICs...).

Conclusions

- Reducing precision can free resources to increase resolution of weather and climate models.
- Single precision is providing almost identical forecast skill when compared to double precision simulations.
- For single precision, savings are mainly generated via a reduction of cash misses and improved vectorization.
- A further reduction beyond single precision for expensive kernels is possible and promising.
- Verificarlo can be used to test for the impact of a reduction in precision.
- We will need better performance models to drive precision reduction in the future.

Funding

Funded by the
European Union

The presenter gratefully acknowledges the funding a Royal Society University Research Fellowship and from the ESIWACE and ESIWACE2 project. The ESIWACE and ESIWACE2 projects have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 675191 and 823988.