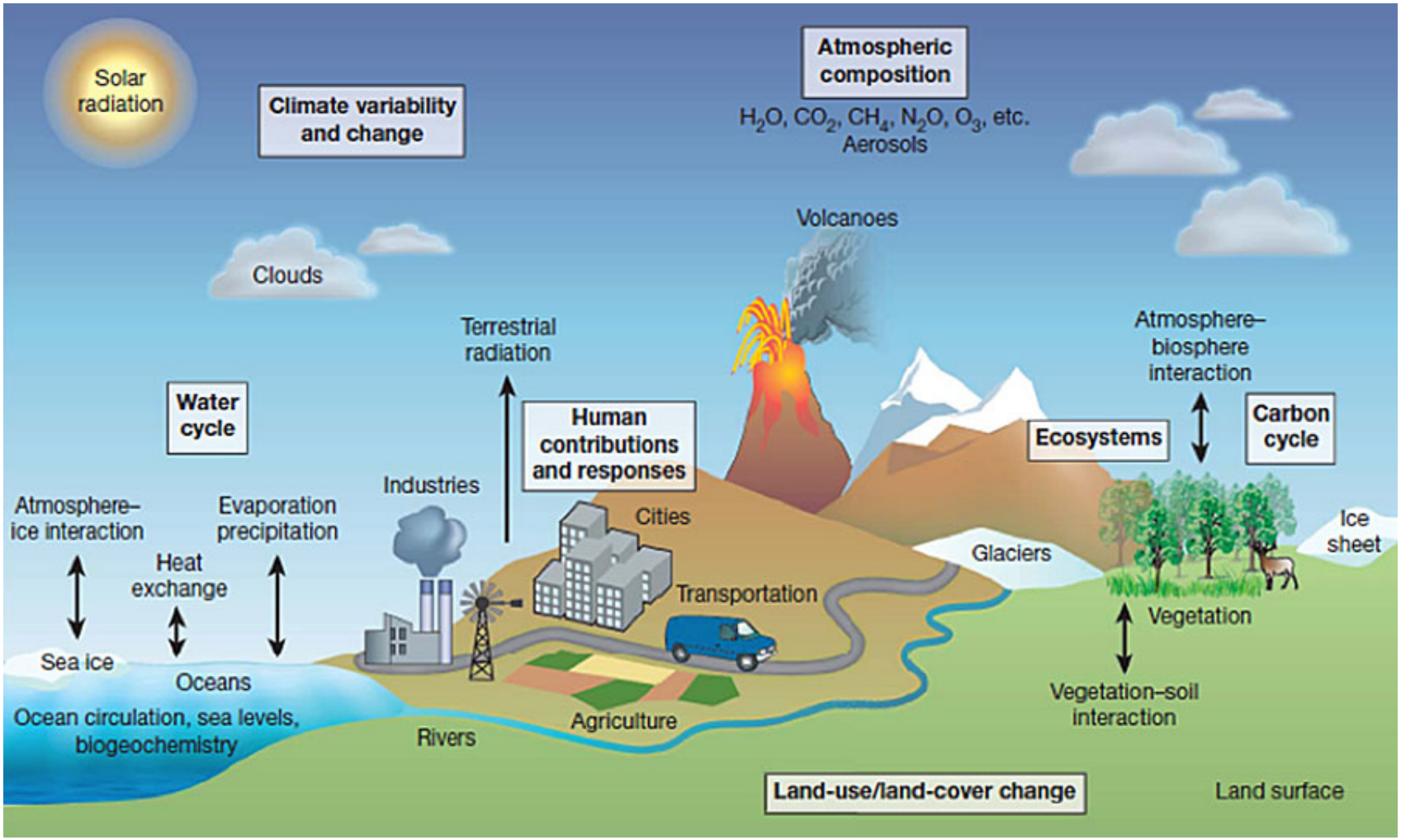
Reduced Precision Computing for Earth System Modelling

Peter Düben

European Centre for Medium-Range Weather Forecasts (ECMWF)

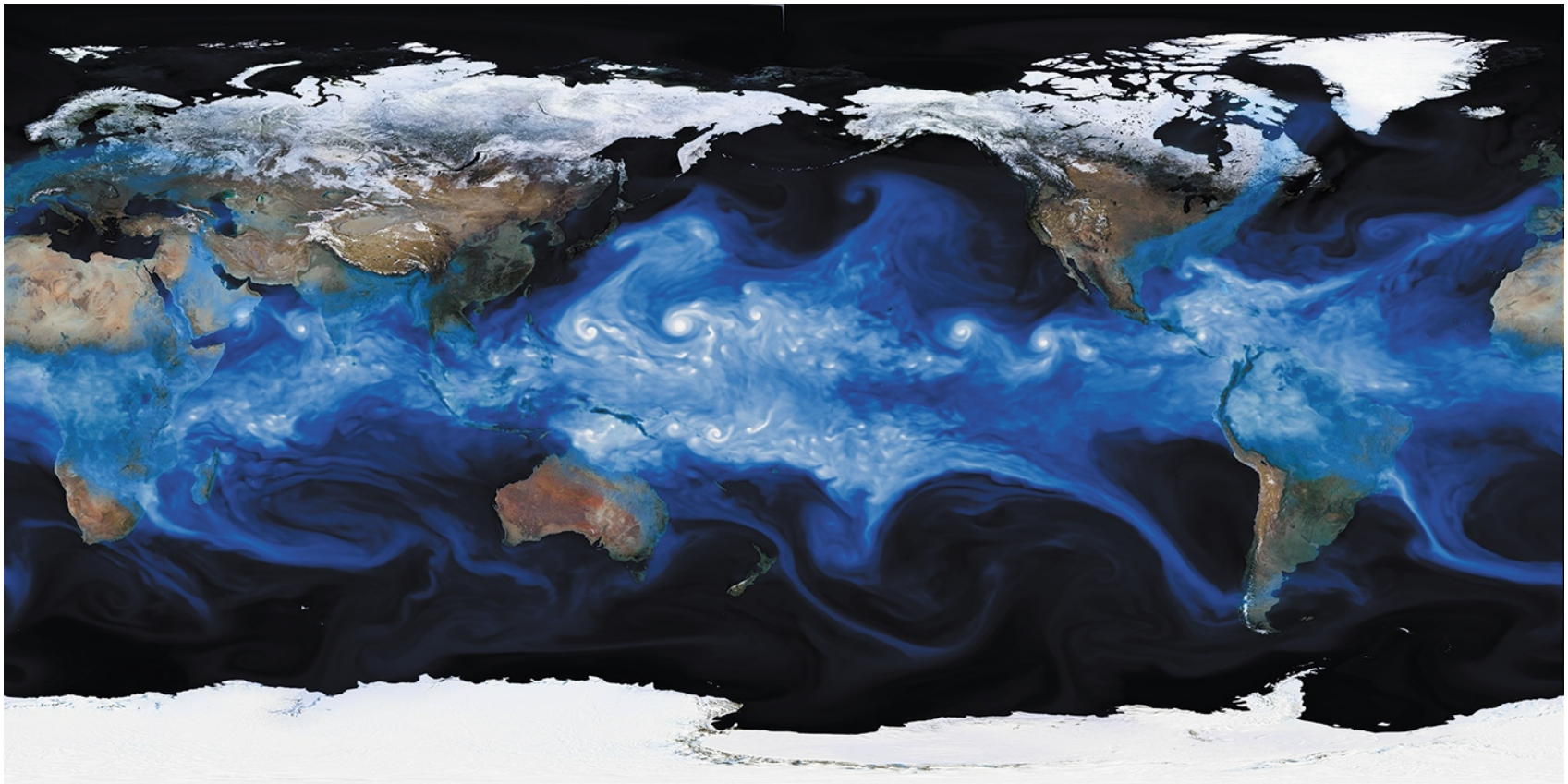
Predicting weather and climate: Why is it so hard?

Predicting weather and climate: Why is it so hard?



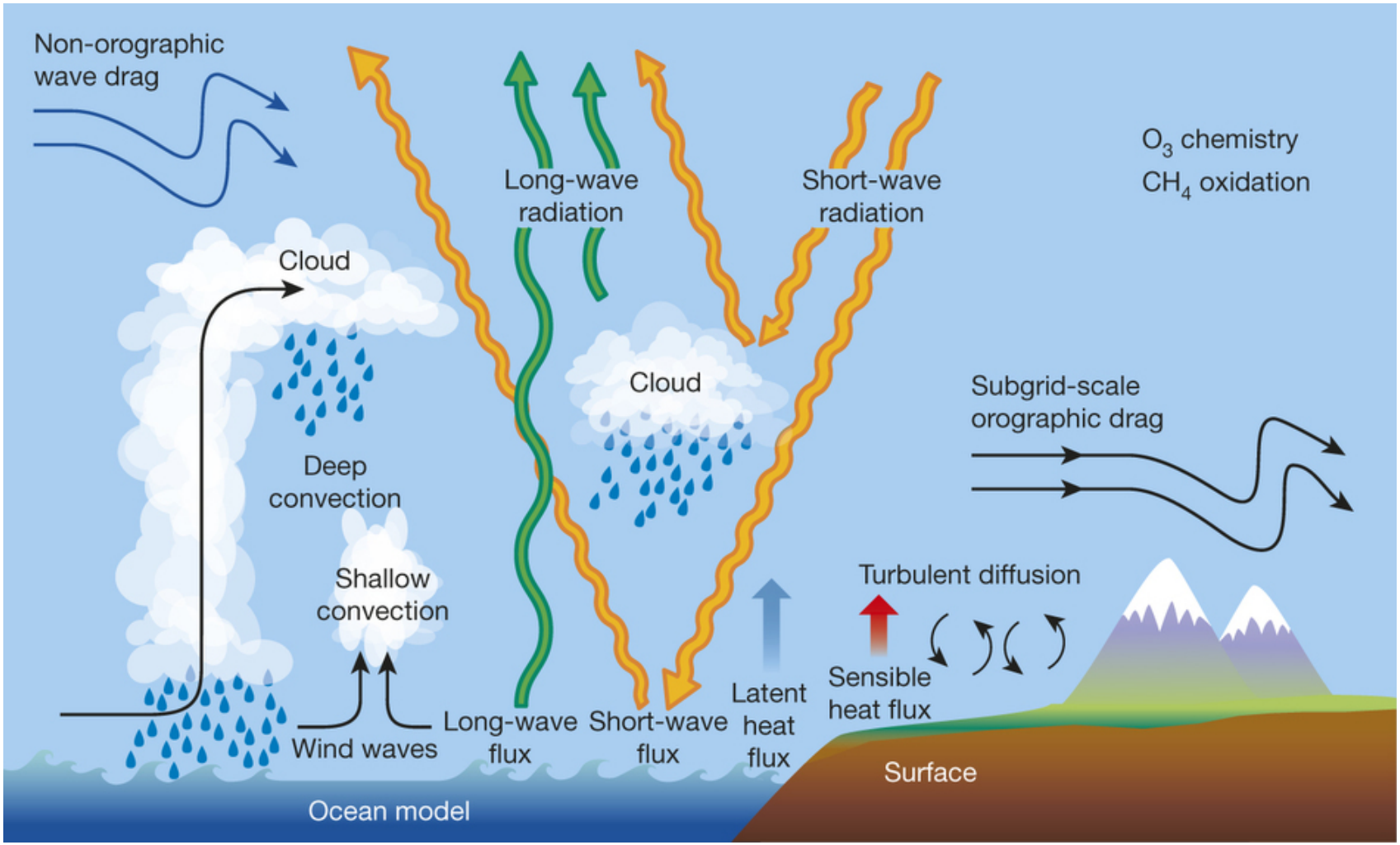
www.gfdl.noaa.gov

Predicting weather and climate: Why is it so hard?



Michael Wehner and Prabhat

Predicting weather and climate: Why is it so hard?



Bauer et al. Nature 2015

Predicting weather and climate: Why is it so hard?

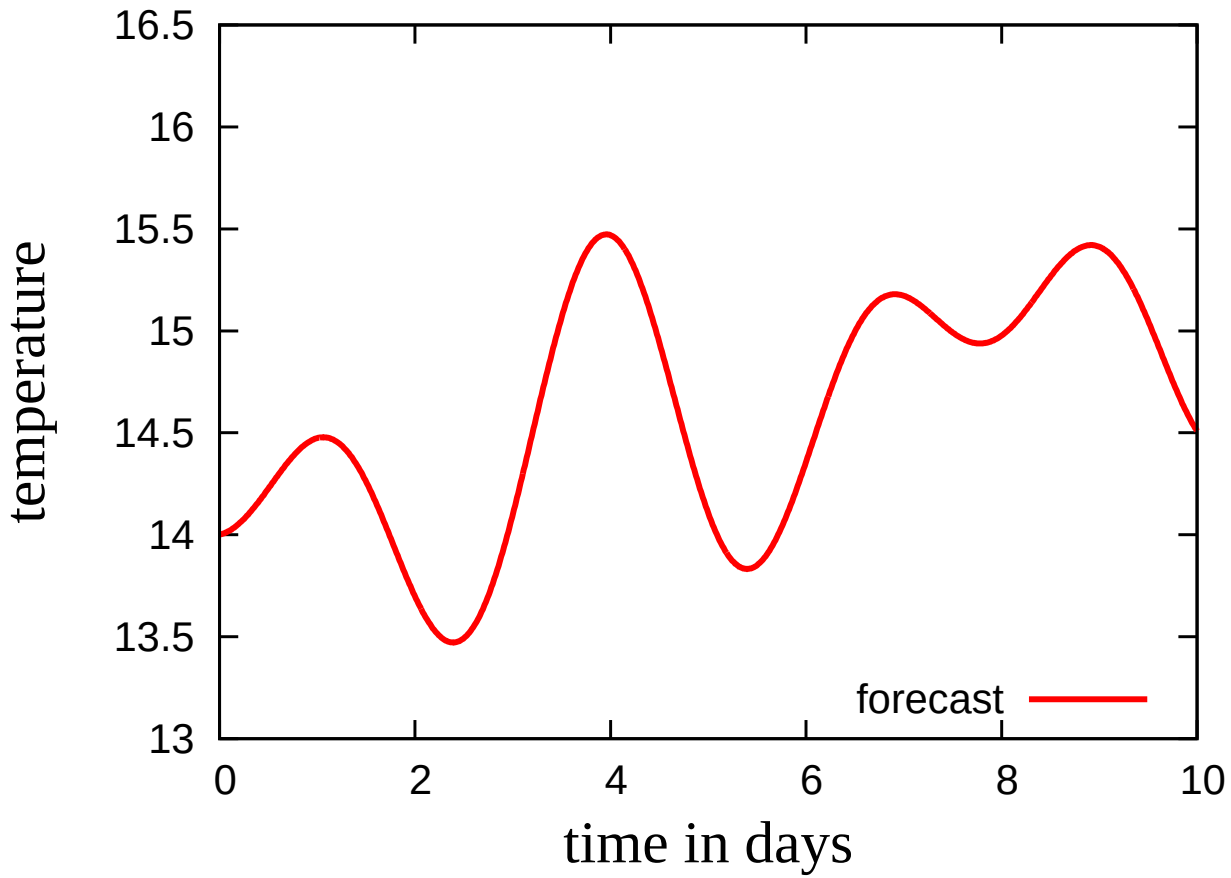
National Geographic Creative

Predicting weather and climate: Why is it so hard?

National Geographic Creative

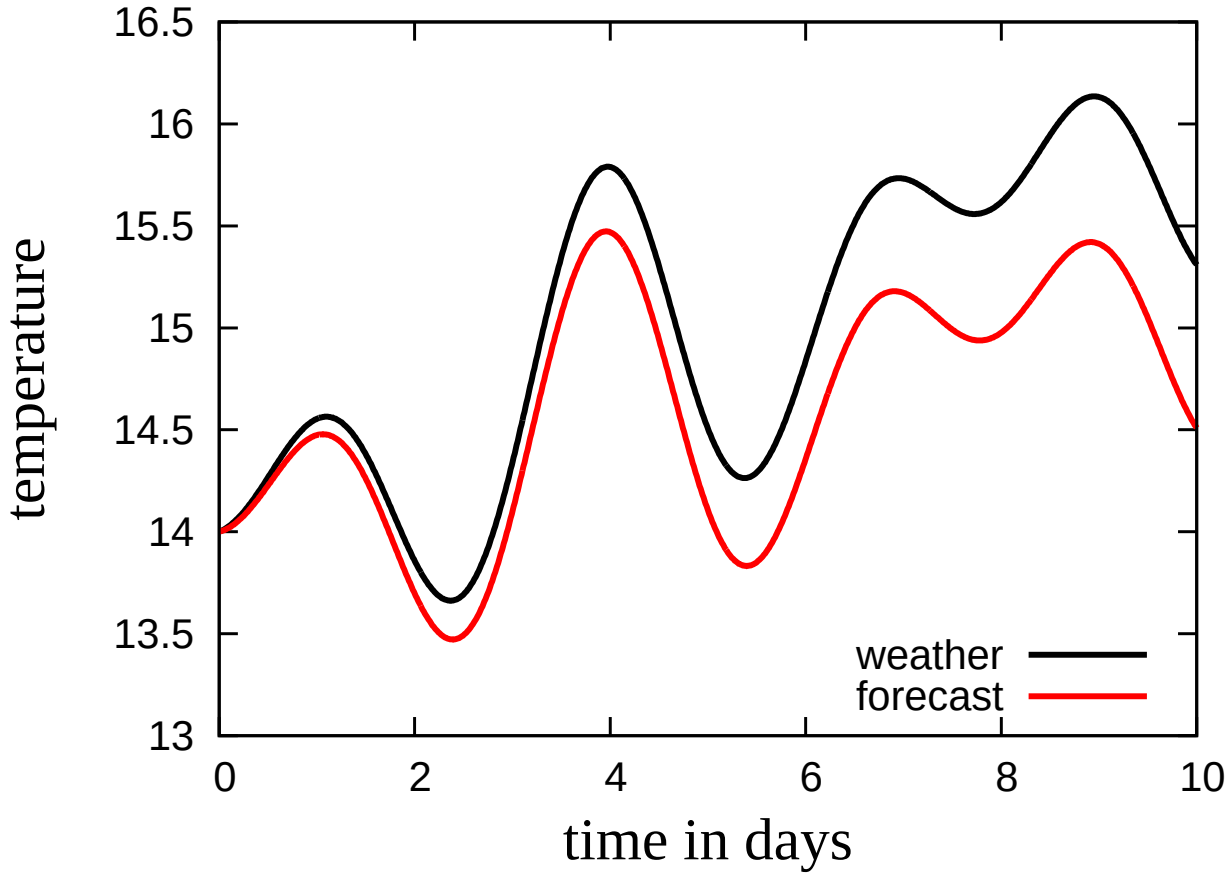
The Earth System is complex, huge and chaotic and we do not have sufficient resolution to resolve all important processes.

How do we treat uncertainties in weather forecasts?



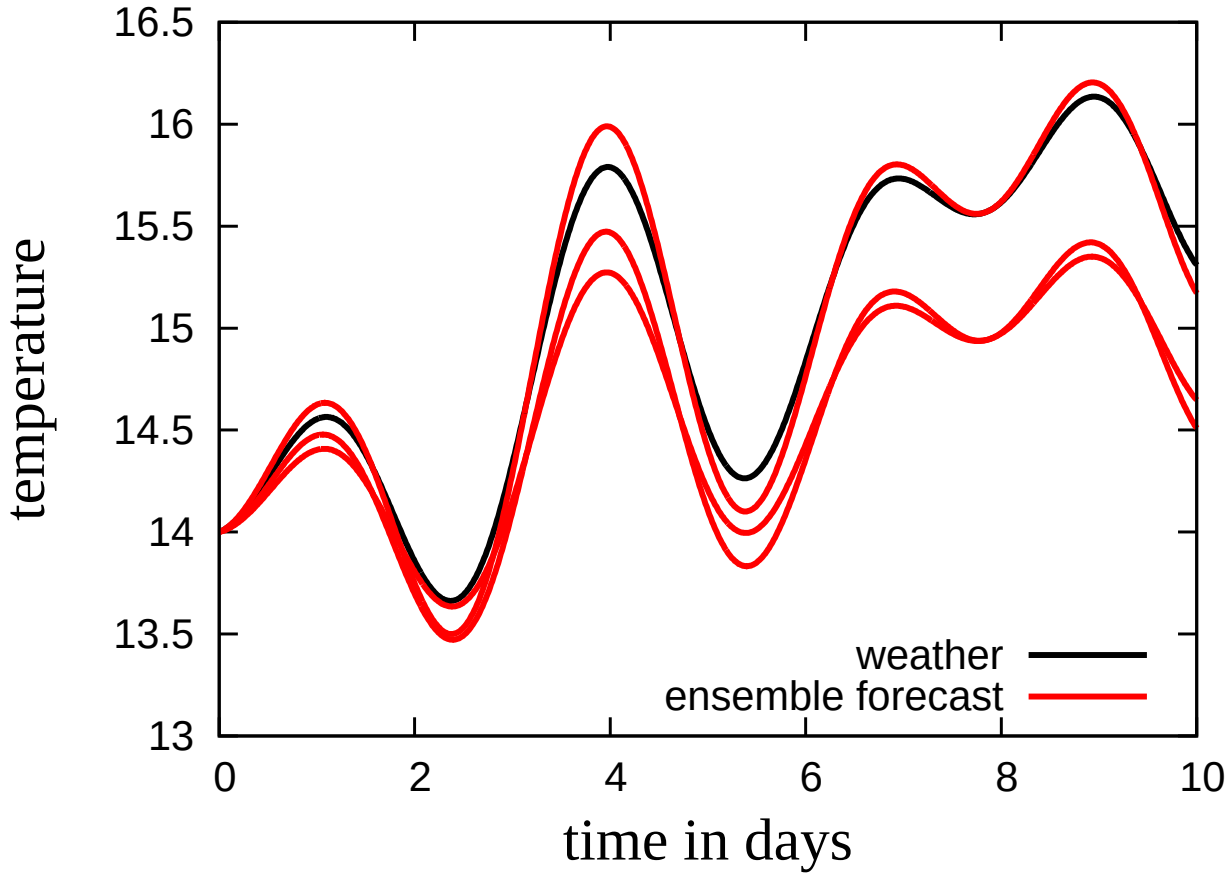
How do we know if we are wrong?

How do we treat uncertainties in weather forecasts?



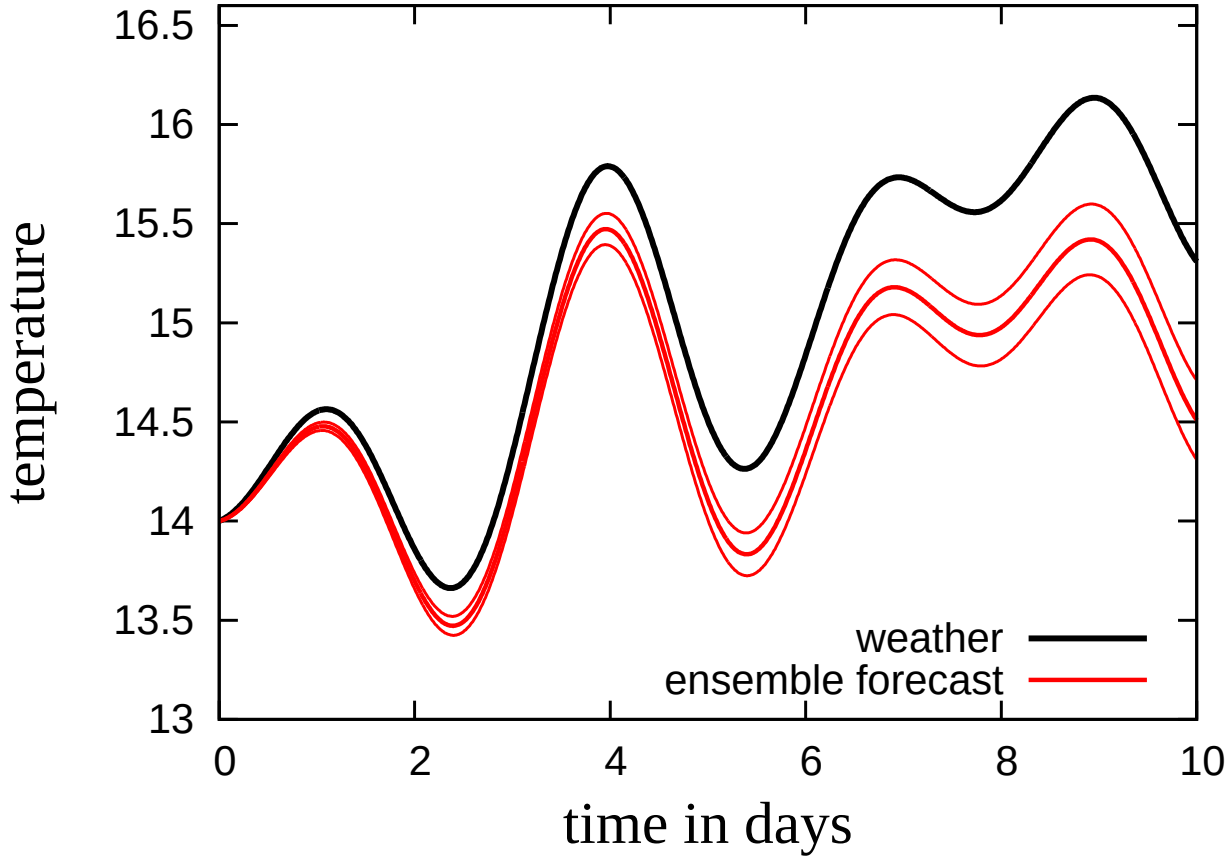
How do we know if we are wrong?

How do we treat uncertainties in weather forecasts?



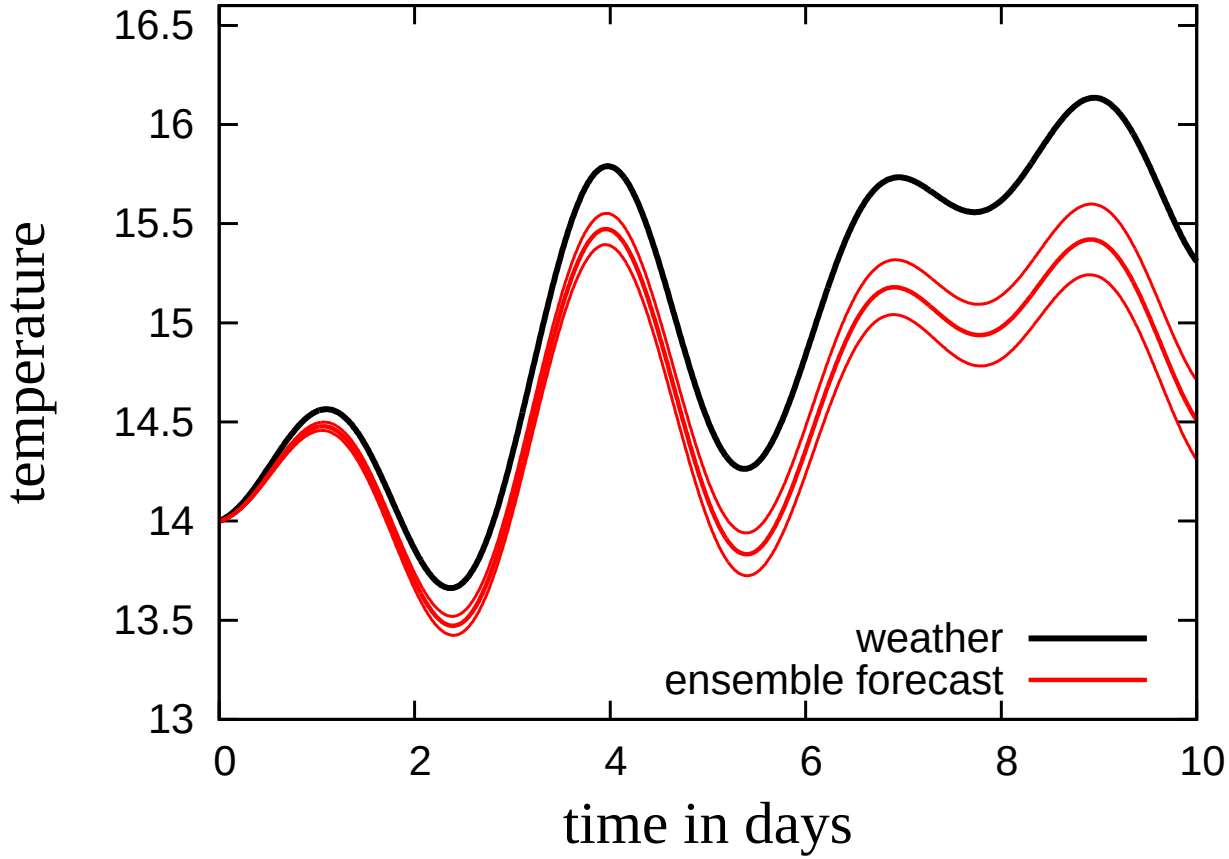
The ensemble spread holds information about forecast uncertainty.

How do we treat uncertainties in weather forecasts?



Ensemble forecasts can go wrong.

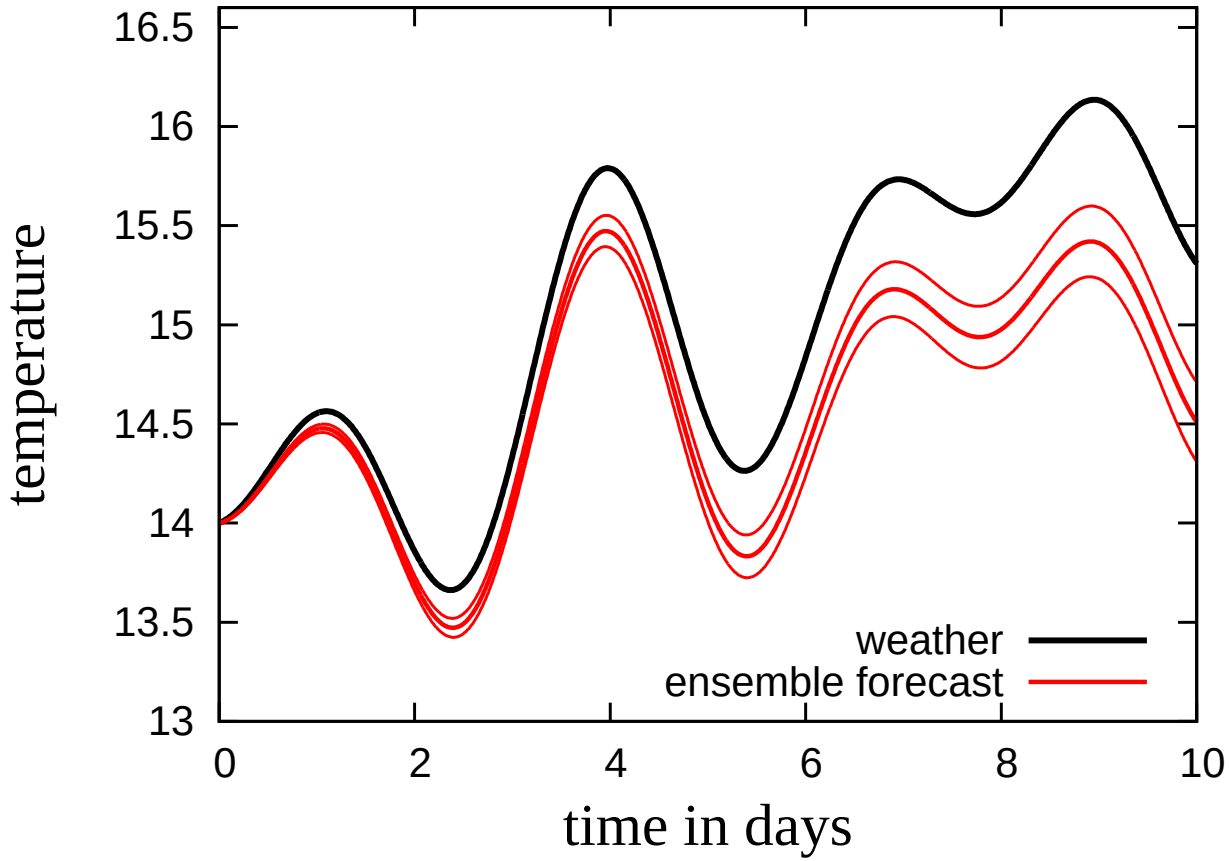
How do we treat uncertainties in weather forecasts?



Ensemble forecasts can go wrong.

We introduce stochastic parametrisation schemes and perturbations to initial conditions to improve ensemble spread.

How do we treat uncertainties in weather forecasts?

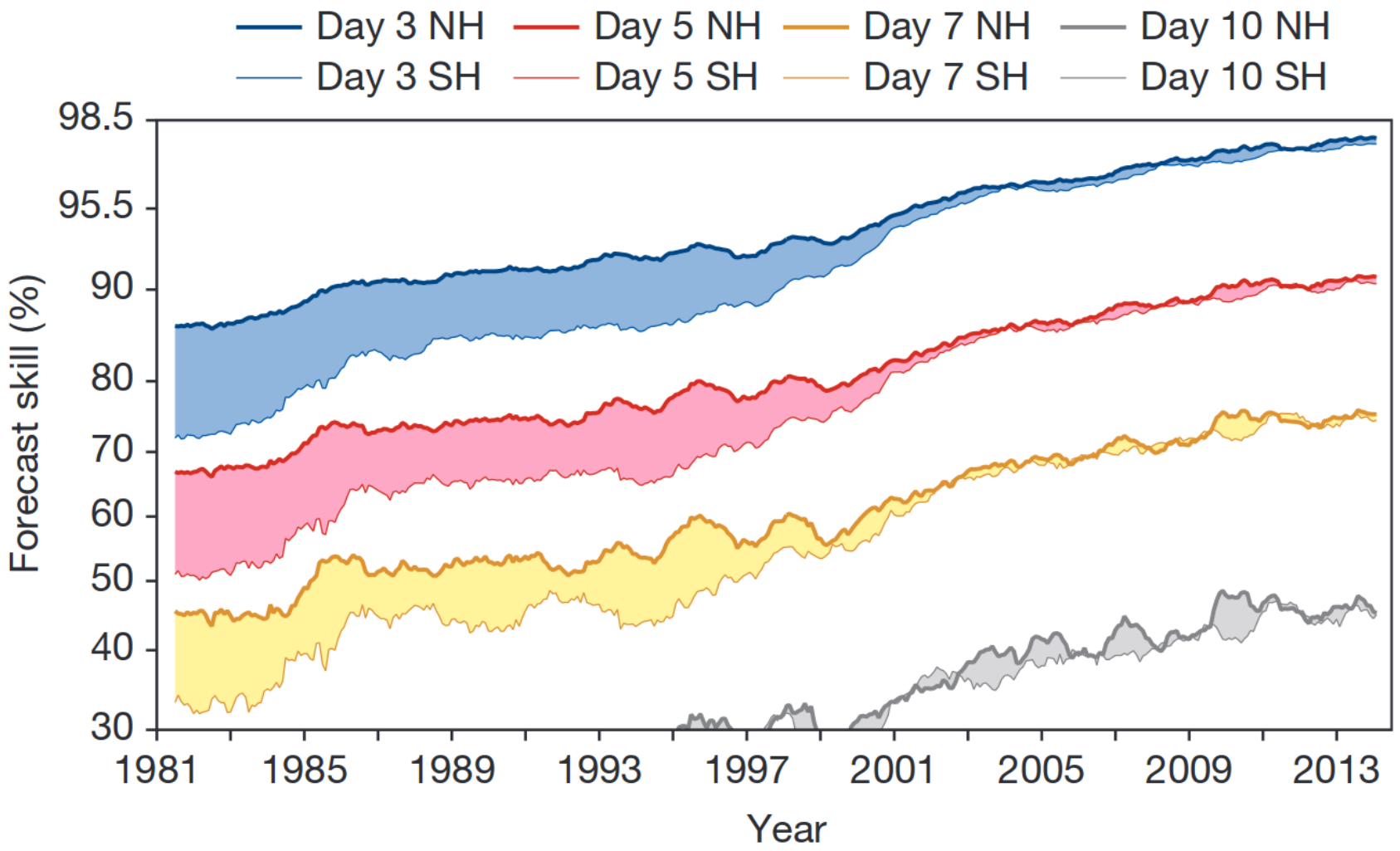


Ensemble forecasts can go wrong.

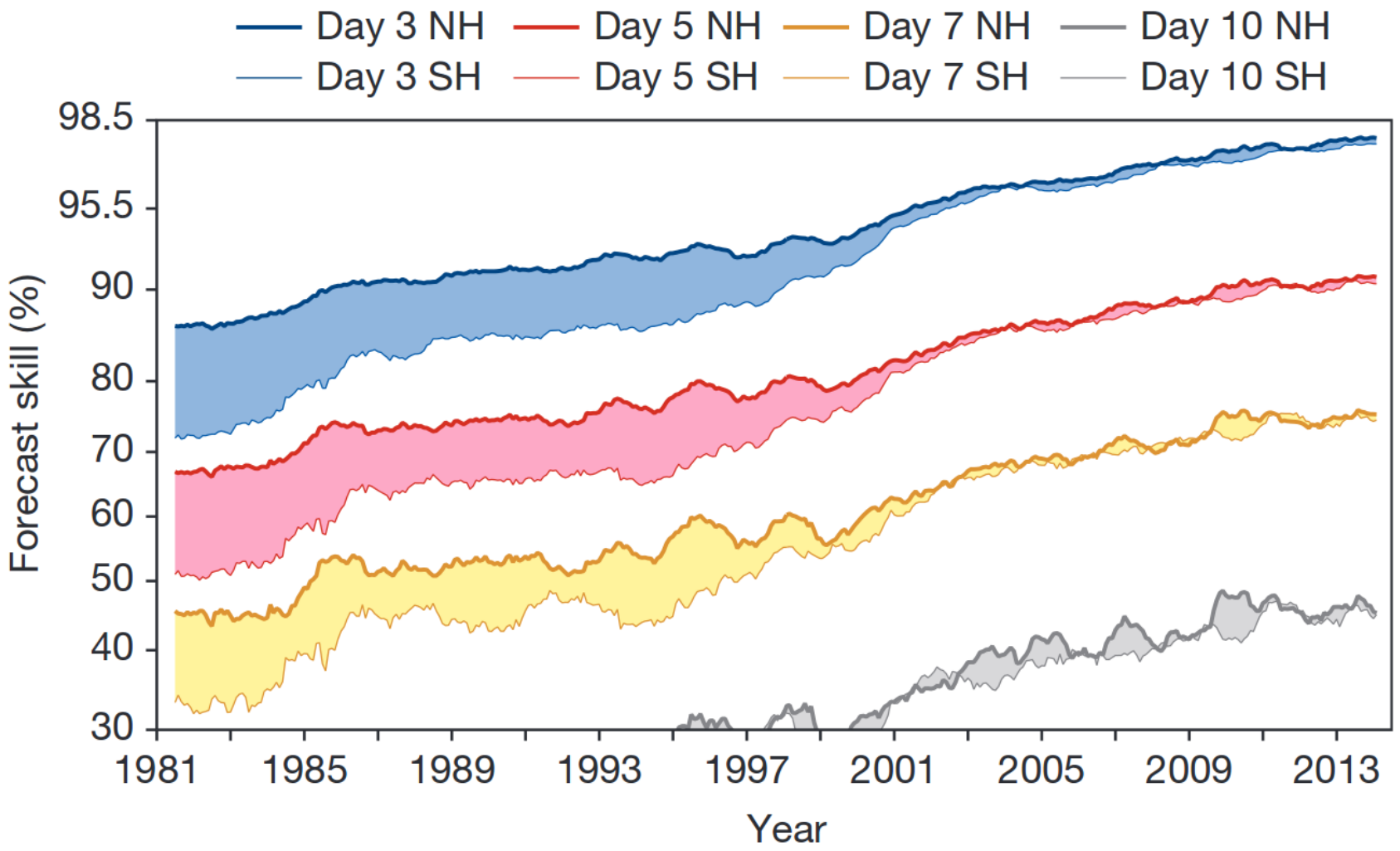
We introduce stochastic parametrisation schemes and perturbations to initial conditions to improve ensemble spread.

These schemes are typically “local” and lack physical justification.

Forecast skill is still improving



Forecast skill is still improving



Higher resolution in weather models → improved forecast skill.

The future of High Performance Computing

- Weather and climate models are high performance computing applications.
- Individual processors will not be faster.
  $\rightarrow$ Parallelisation ($> 10^6$ parallel processing units).
- Power consumption will be a big problem.
- Scalability and performance will influence decisions in model development.

The future of High Performance Computing

- Weather and climate models are high performance computing applications.
- Individual processors will not be faster.
  $\rightarrow$ Parallelisation ($> 10^6$ parallel processing units).
- Power consumption will be a big problem.
- Scalability and performance will influence decisions in model development.

The free lunch is over.

Less numerical precision $\rightarrow$ 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 Reading:
double precision (64 bits): 14.5611925125122 07°C
single precision (32 bits): 14.5611925°C
half precision (16 bits): 14.5625°C

A short introduction to bit representation

- The computer represents an integer number as a string of 32 bits. Each bit represents a power of two:

$$102090 = 0\cdot 2^0 + 1\cdot 2^1 + 0\cdot 2^2 + 1\cdot 2^3 + 0\cdot 2^4 + 0\cdot 2^5 + 1\cdot 2^6 \ldots = \sum_{i=0}^{31} b_i 2^i$$

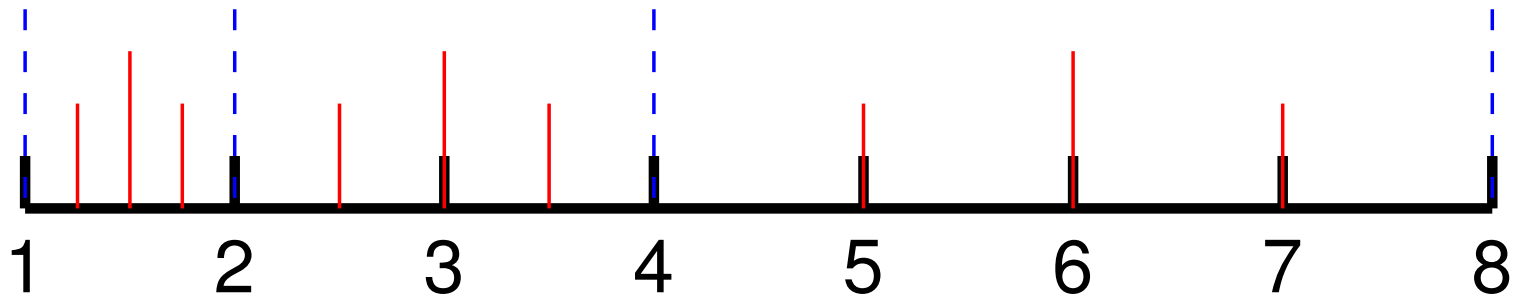
A short introduction to bit representation

- The computer represents an integer number as a string of 32 bits. Each bit represents a power of two:

$$102090 = 0\cdot 2^0 + 1\cdot 2^1 + 0\cdot 2^2 + 1\cdot 2^3 + 0\cdot 2^4 + 0\cdot 2^5 + 1\cdot 2^6 \ldots = \sum_{i=0}^{31} b_i 2^i$$

- A real number a is represented as a 64 bit floating point number:

$$a = (-1)^S \left(1 + \sum_{i=1}^{52} b_{-i} 2^{-i}\right) 2^E, \quad \text{where} \quad E = \left(\sum_{i=0}^{10} e_i 2^i\right) - 1023.$$



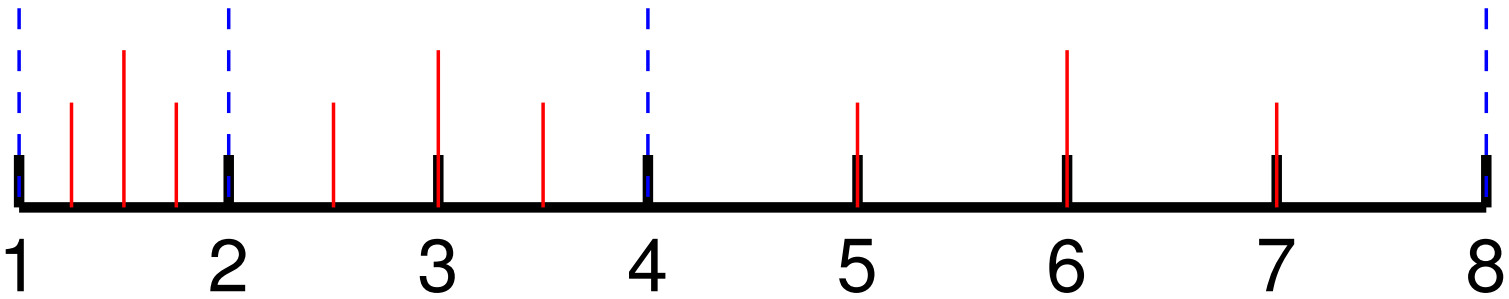
A short introduction to bit representation

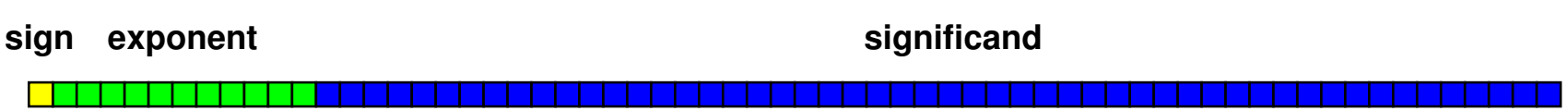
- The computer represents an integer number as a string of 32 bits. Each bit represents a power of two:

$$102090 = 0\cdot 2^0 + 1\cdot 2^1 + 0\cdot 2^2 + 1\cdot 2^3 + 0\cdot 2^4 + 0\cdot 2^5 + 1\cdot 2^6 \ldots = \sum_{i=0}^{31} b_i 2^i$$

- A real number a is represented as a 64 bit floating point number:

$$a = (-1)^S \left(1 + \sum_{i=1}^{52} b_{-i} 2^{-i}\right) 2^E, \quad \text{where} \quad E = \left(\sum_{i=0}^{10} e_i 2^i\right) - 1023.$$

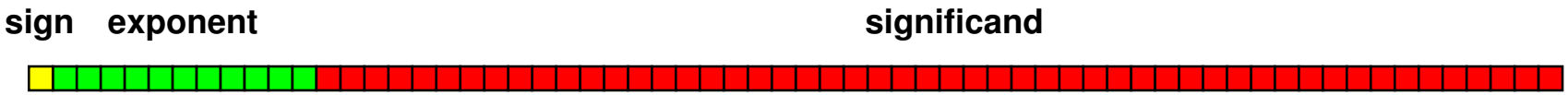




Approaches to inexact floating point units

Stochastic processor

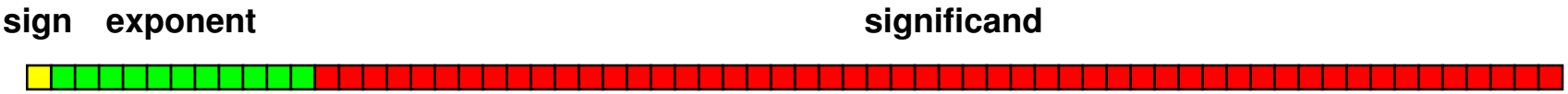
- If we reduce the applied voltage or the wall clock time beyond a certain level, we will get hardware errors, but we will save power.
- The error rate of a stochastic processor can be reduced massively, if the architecture is changed.



Approaches to inexact floating point units

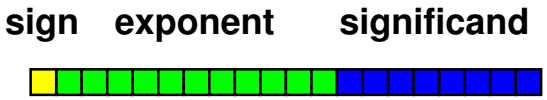
Stochastic processor

- If we reduce the applied voltage or the wall clock time beyond a certain level, we will get hardware errors, but we will save power.
- The error rate of a stochastic processor can be reduced massively, if the architecture is changed.



Pruning

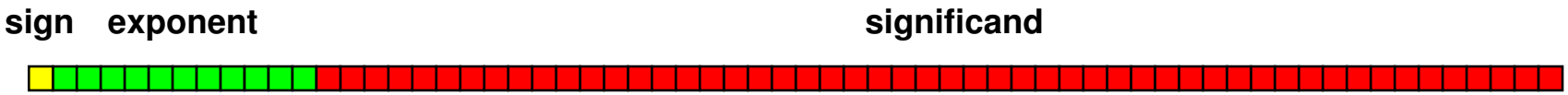
Parts of the CPU that are hardly used or do not have a strong influence on significant bits are removed.



Approaches to inexact floating point units

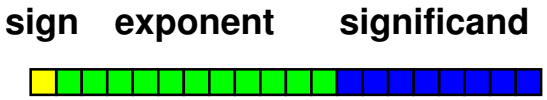
Stochastic processor

- If we reduce the applied voltage or the wall clock time beyond a certain level, we will get hardware errors, but we will save power.
- The error rate of a stochastic processor can be reduced massively, if the architecture is changed.



Pruning

Parts of the CPU that are hardly used or do not have a strong influence on significant bits are removed.



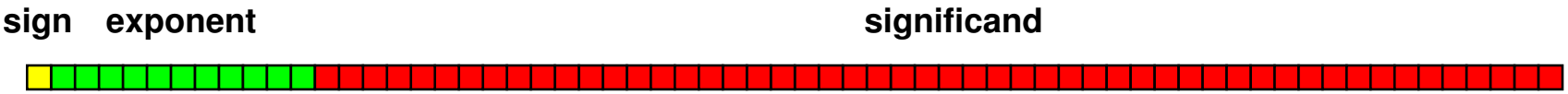
Field Programmable Gate Array (FPGA)

- FPGAs are integrated circuits that can be configured by the user.
- Numerical precision can be customised to the application.

Approaches to inexact floating point units

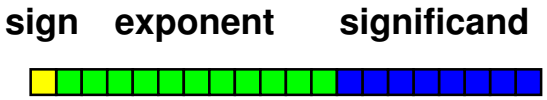
Stochastic processor

- If we reduce the applied voltage or the wall clock time beyond a certain level, we will get hardware errors, but we will save power.
- The error rate of a stochastic processor can be reduced massively, if the architecture is changed.



Pruning

Parts of the CPU that are hardly used or do not have a strong influence on significant bits are removed.



Field Programmable Gate Array (FPGA)

- FPGAs are integrated circuits that can be configured by the user.
- Numerical precision can be customised to the application.

Easiest way: double $\rightarrow$ single $\rightarrow$ half.

Why should we use reduced precision in weather and climate predictions?

This is what we want to represent in an atmosphere model.

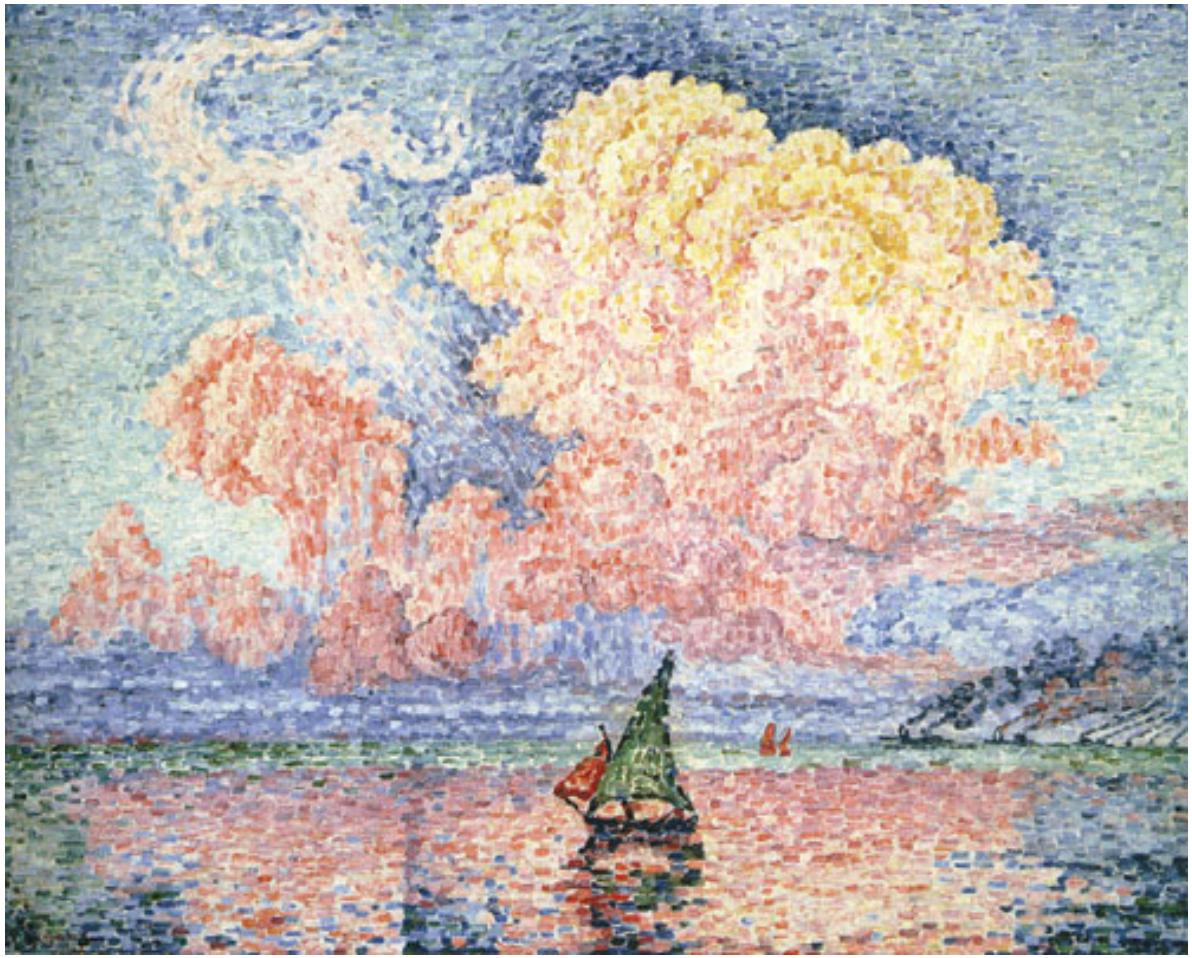
Why should we use reduced precision in weather and climate predictions?

This is how we represent the atmosphere.

Steve Greaves

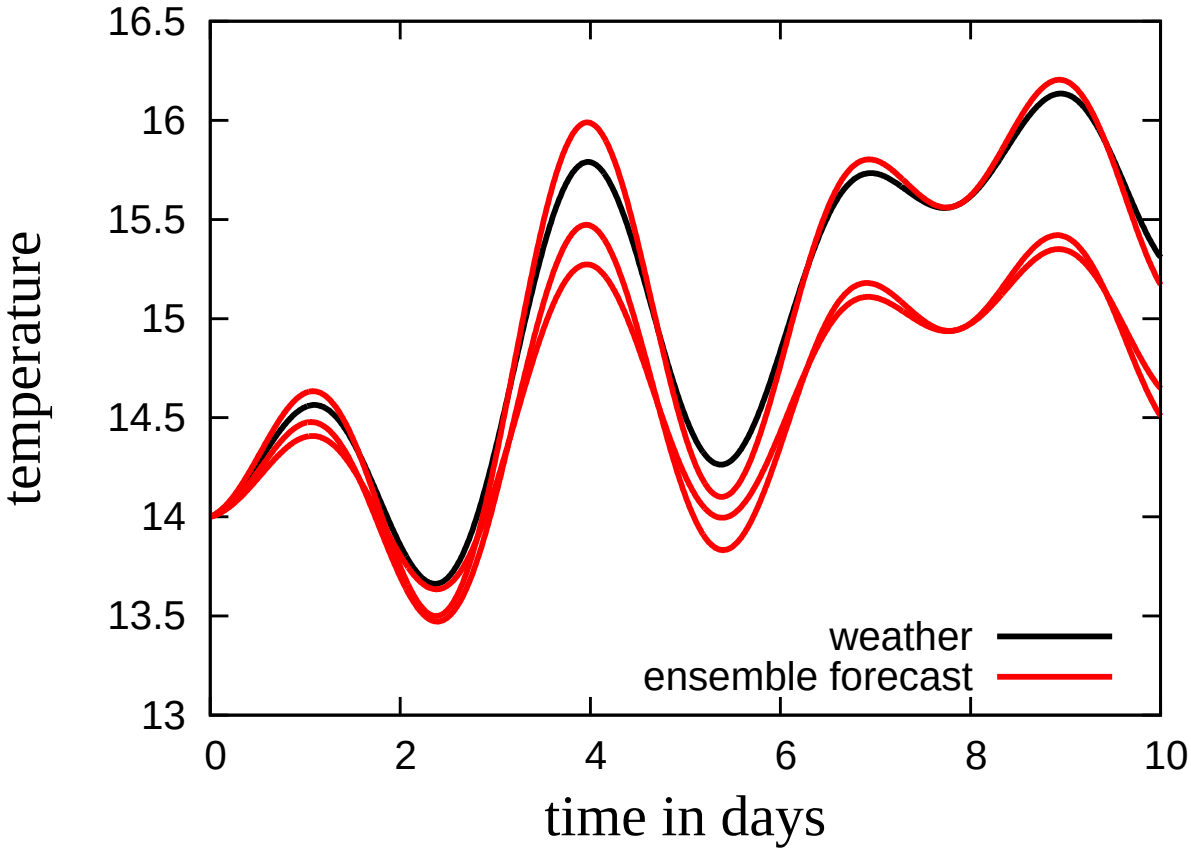
Why should we use reduced precision in weather and climate predictions?

Can we represent the atmosphere like this?



Antibes 1916

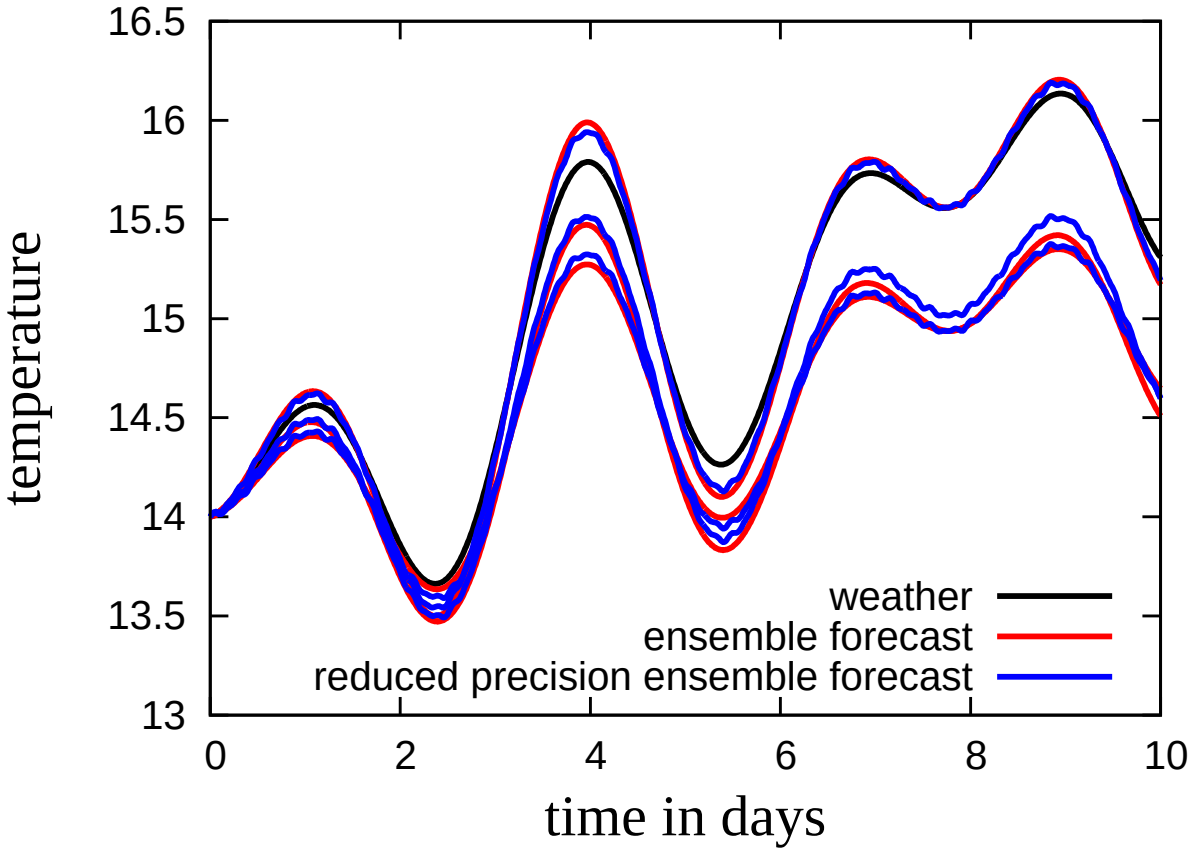
Two research questions



Will our models fail if we reduce precision?

Can we identify the optimal level of precision?

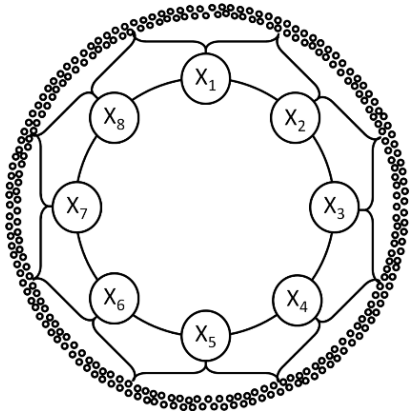
Two research questions



Will our models fail if we reduce precision?

Can we identify the optimal level of precision?

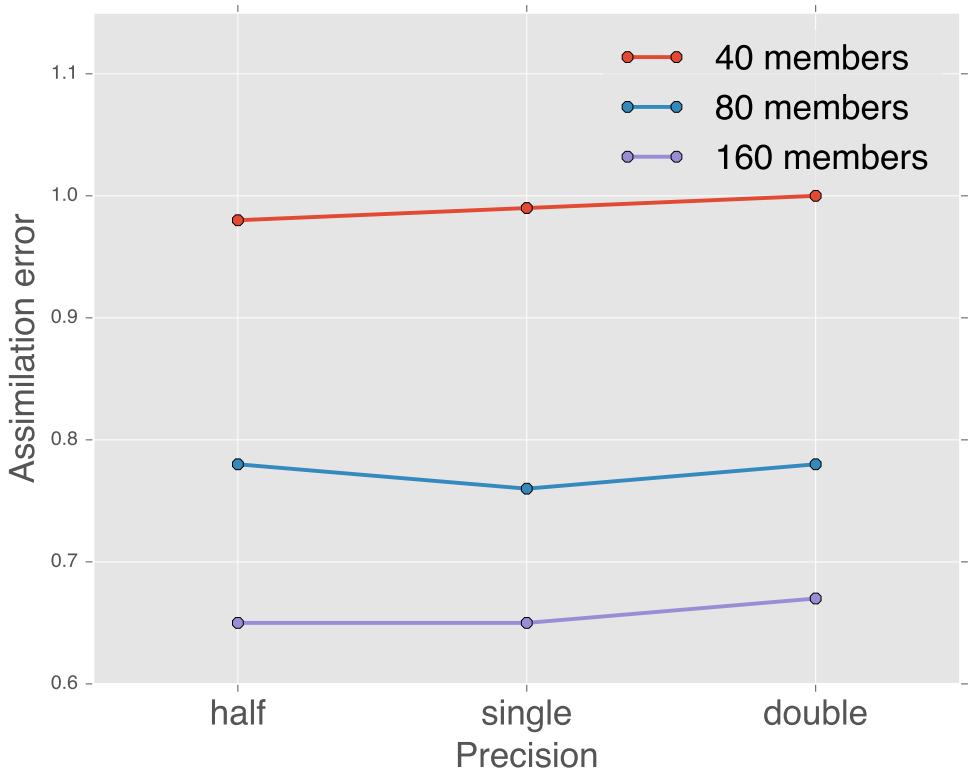
Lorenz ’96 on FPGAs



- We implemented the Lorenz ’96 model on FPGAs in cooperation with the group of Wayne Luk at Imperial College.
- We scale the size of the Lorenz model to the size of a high performance application with more than 100 million degrees-of-freedom.
- Simulations with reduced precision (17 bits for X; 14 bits for Y) are more than two times faster compared to simulations in single precision.
- The error in these simulations is comparable to a parameter change of 1%.

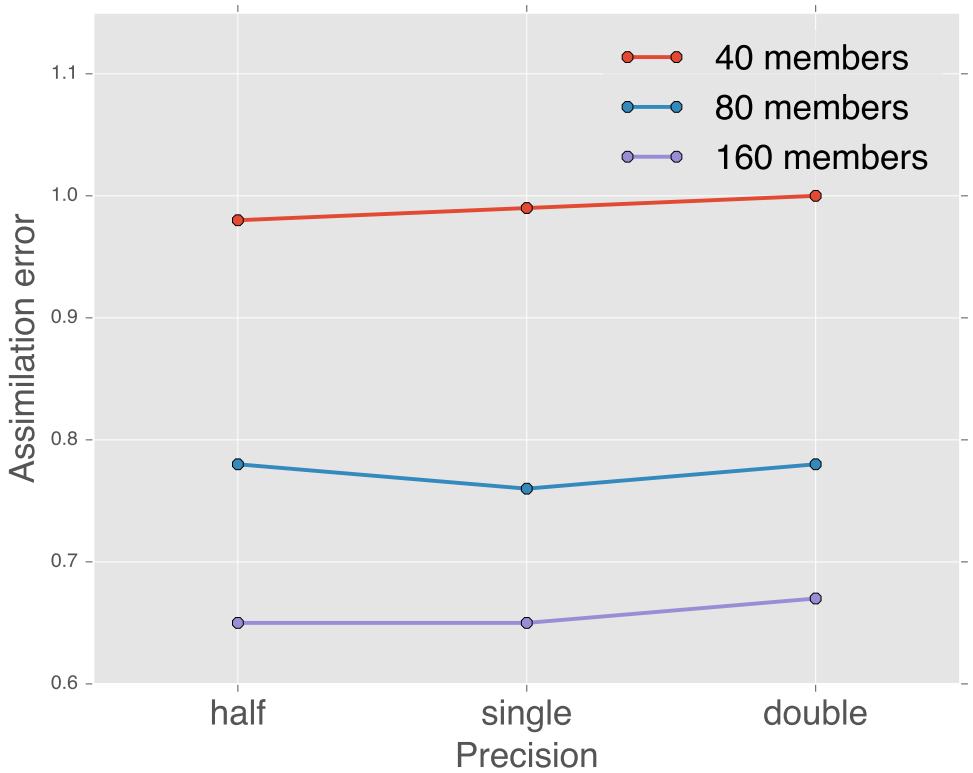
Düben et al. JAMES 2015, Russel et al. FCCM 2015.

Data assimilation with reduced precision
PhD student Samuel Hatfield



Data assimilation in Lorenz'95 using an Ensemble Kalman filter.

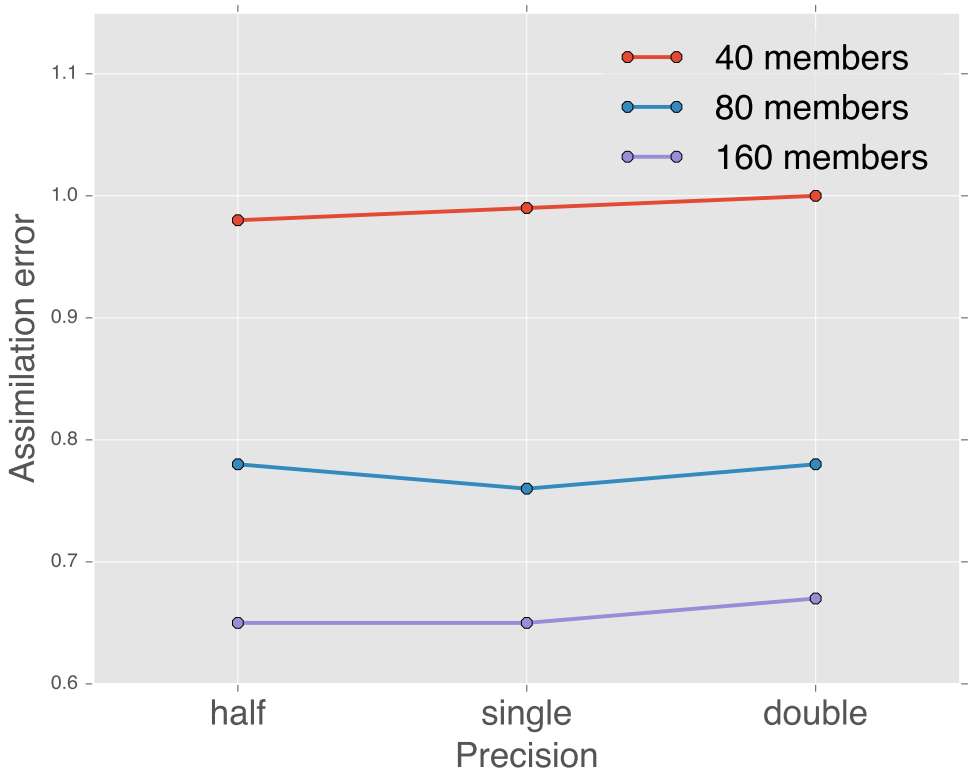
Data assimilation with reduced precision
PhD student Samuel Hatfield



Data assimilation in Lorenz'95 using an Ensemble Kalman filter.

A large ensemble at low precision is better than a small ensemble at high precision at the same computing cost.

Data assimilation with reduced precision
PhD student Samuel Hatfield



Data assimilation in Lorenz'95 using an Ensemble Kalman filter.

A large ensemble at low precision is better than a small ensemble at high precision at the same computing cost.

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

Reduced precision in an atmosphere model

- We calculate weather forecasts with a spectral dynamical core (full 3D dynamics on the globe but no physics).
- Floating point precision is reduced to 20 bits (instead of 64) using an emulator in almost the entire model.
- We estimate savings for reduced precision in cooperation with computer scientists (the groups of Krishna Palem - Rice University, Christian Enz - EPFL and John Augustine - IITM).

Reduced precision in an atmosphere model

Resolution	Precision in number of bits	Normalised Energy Demand	Mean error Z500 at day 2
235 km	64	1.0	2.3
315 km	64	0.47	4.5
235 km	**20**	0.29	2.5

Reduced precision in an atmosphere model

Resolution	Precision in number of bits	Normalised Energy Demand	Mean error Z500 at day 2
235 km	64	1.0	2.3
315 km	64	0.47	4.5
235 km	**20**	0.29	2.5

To save power a reduction in precision is much more efficient when compared to a reduction in resolution.

Reduced precision in an atmosphere model

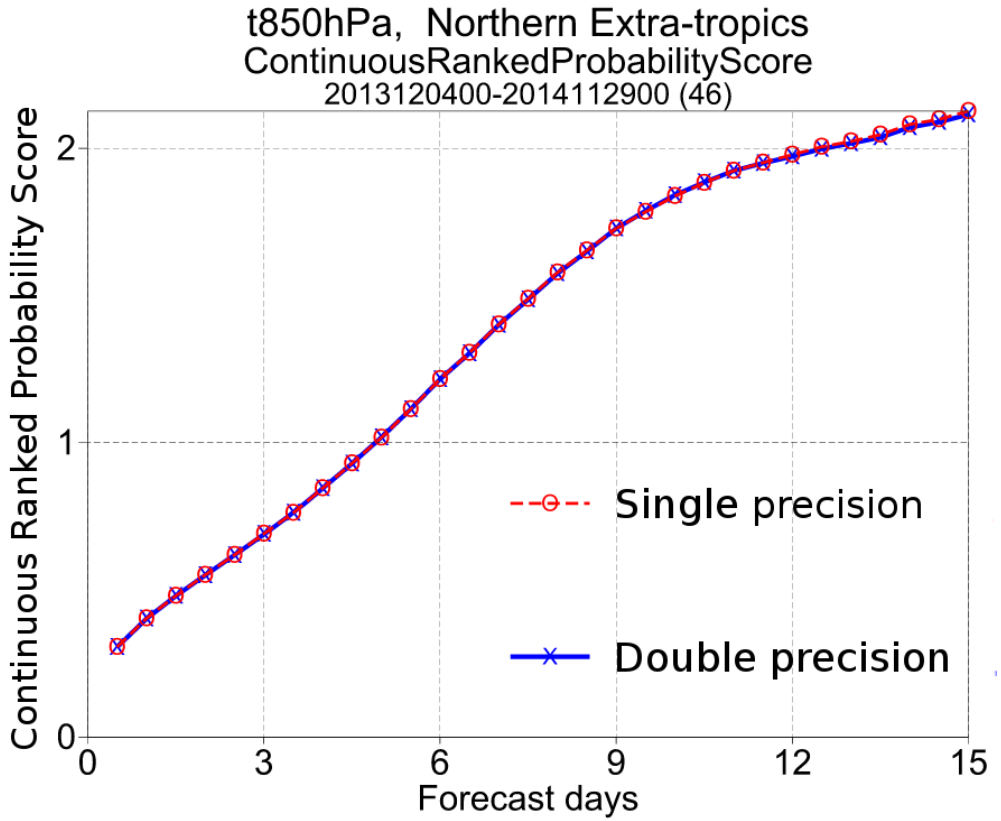
Resolution	Precision in number of bits	Normalised Energy Demand	Mean error Z500 at day 2
235 km	64	1.0	2.3
315 km	64	0.47	4.5
235 km	**20**	0.29	2.5

To save power a reduction in precision is much more efficient when compared to a reduction in resolution.

Studies with programmable hardware (FPGAs) confirm this result.

Düben et al. MWR 2015; Düben et al. DATE 2015; Düben et al. JAMES 2015; Russel, Düben et al. FCCM 2015.

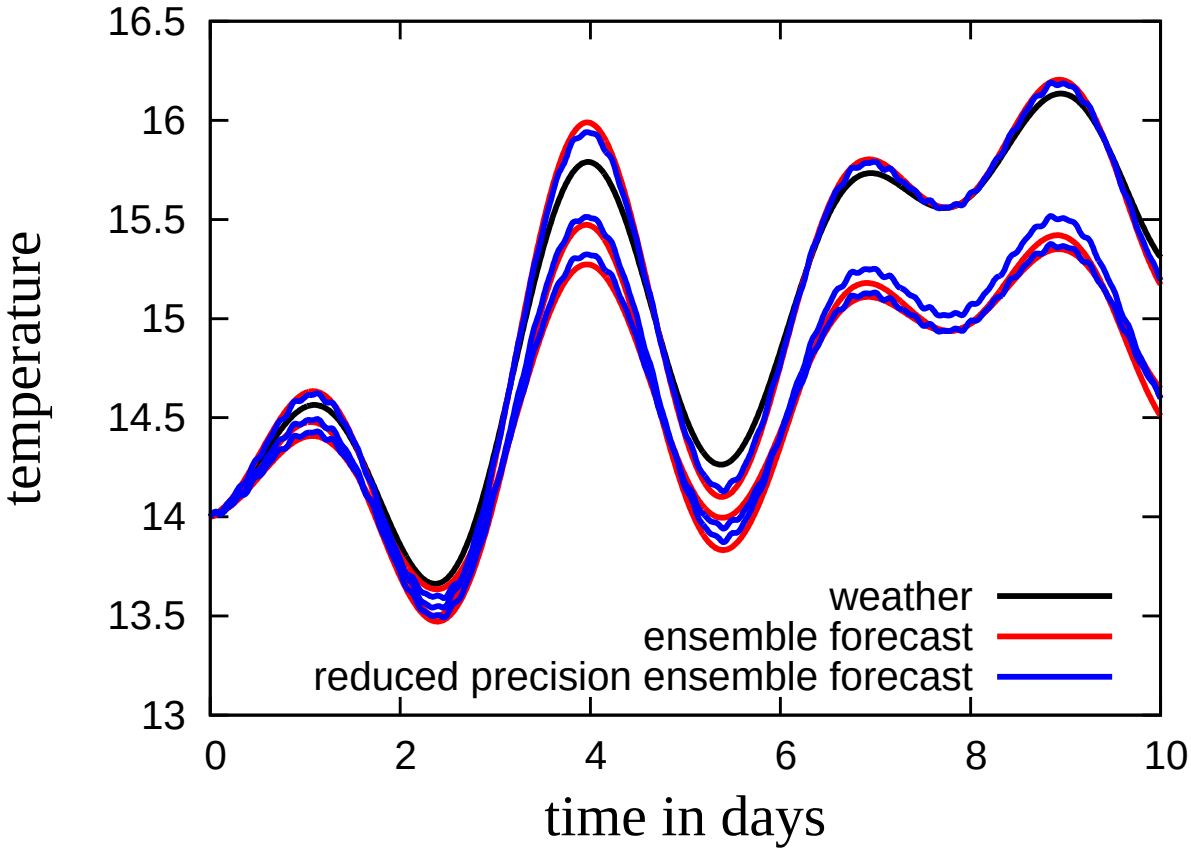
ECMWF's weather forecast model in single precision



- Ensemble forecasts and long-term simulations in double and single precision are almost identical.
- 40% speed-up.
- Single precision for global simulations at 2.5 and 1.0 km resolution.

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

Two research questions



Will our models fail if we reduce precision? - No!

Can we identify the optimal level of precision?

Rounding errors adjusted to model error

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

Rounding errors adjusted to model error

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

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

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

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 data volume by a factor of two.

Rounding errors adjusted to model error

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.

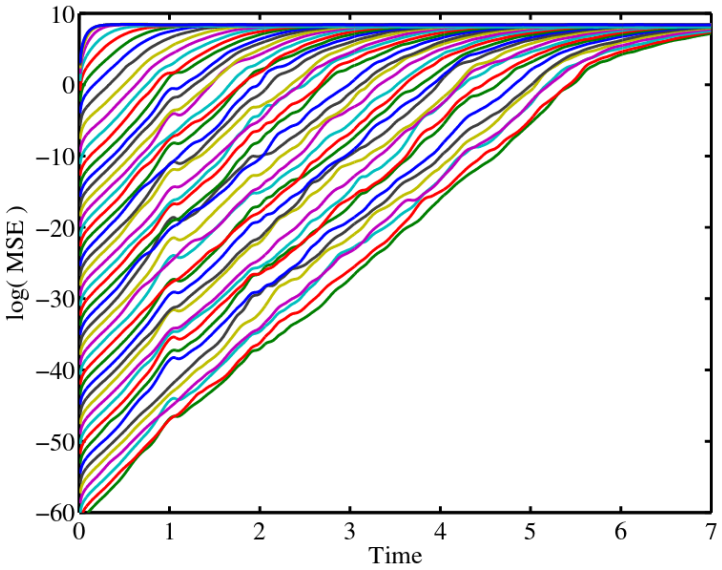
Rounding errors will decrease exponentially with the number of bits.

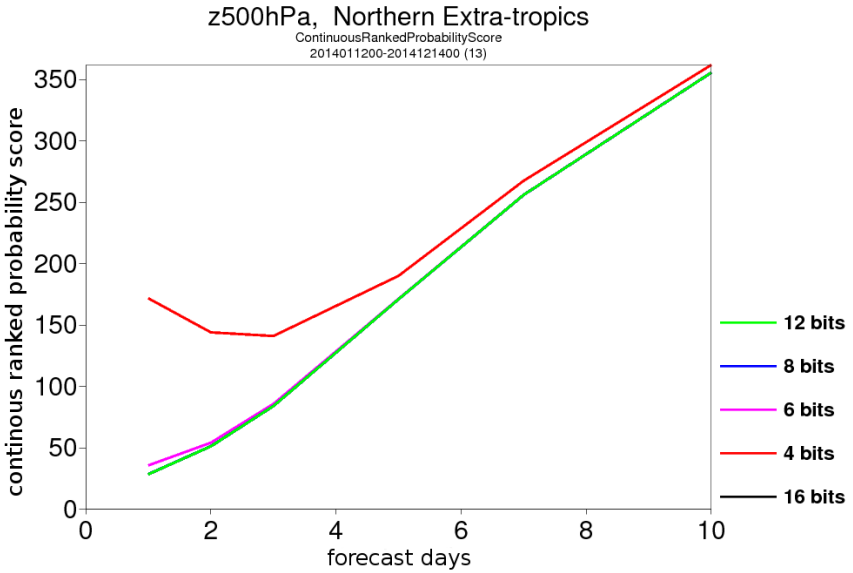
$\rightarrow$ Precision should be reduced linearly with forecast lead time proportional to the leading Lyapunov exponent.

This would reduce data volume by a factor of two.

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

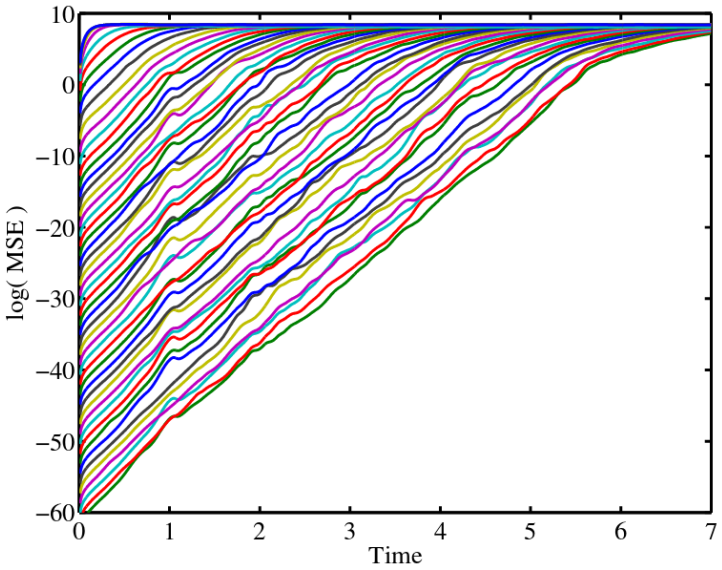
Rounding errors adjusted to model error

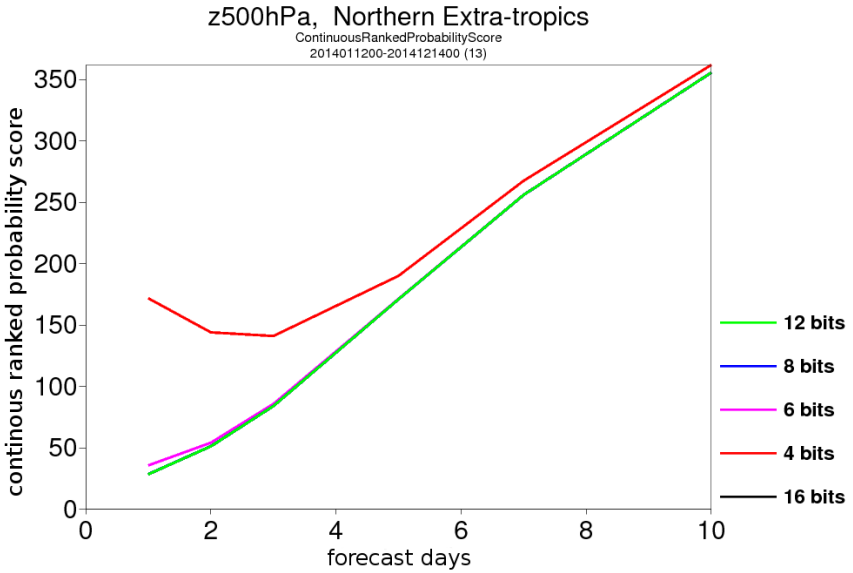




Mean Square Error of simulations with Lorenz'95 and Continuous Ranked Probability Skill for data of ECMWF's ensemble forecasts.

Rounding errors adjusted to model error

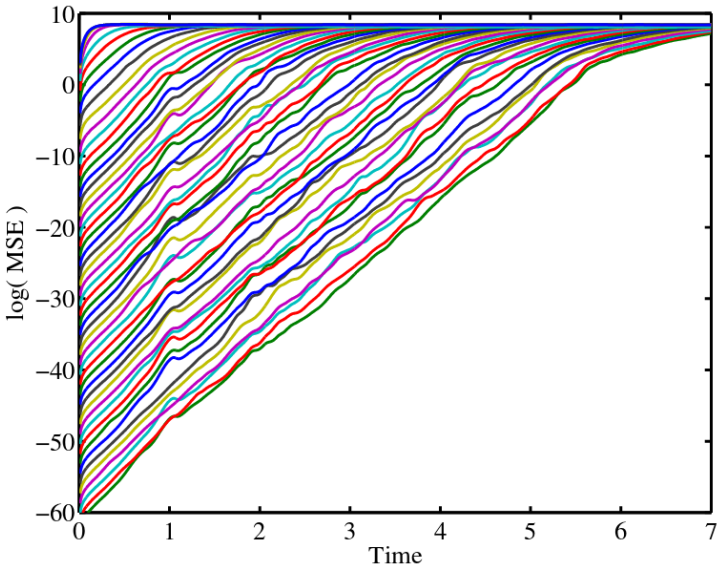


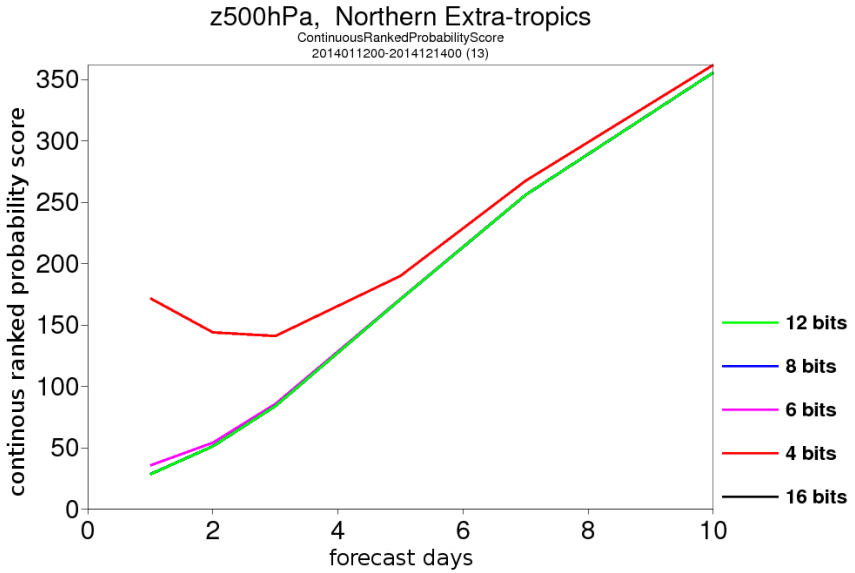


Mean Square Error of simulations with Lorenz'95 and Continuous Ranked Probability Skill for data of ECMWF's ensemble forecasts.

Rounding errors are clearly linked to model error.

Rounding errors adjusted to model error





Mean Square Error of simulations with Lorenz'95 and Continuous Ranked Probability Skill for data of ECMWF's ensemble forecasts.

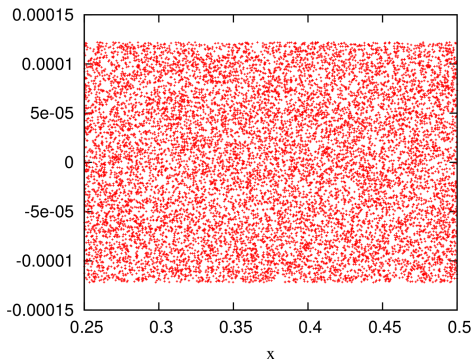
Rounding errors are clearly linked to model error.

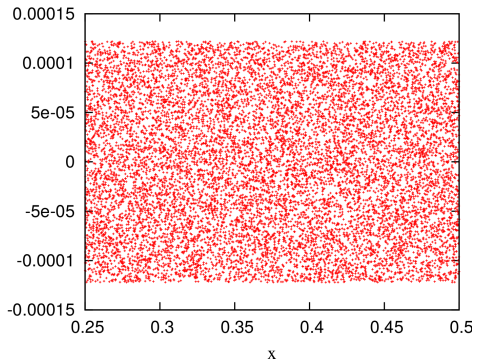
Promising! However, more tests are needed.

Düben et al. JAMES 2015, Cooper, Düben et al. in prep. for MWR

Compare rounding errors with stochastic parametrisation schemes

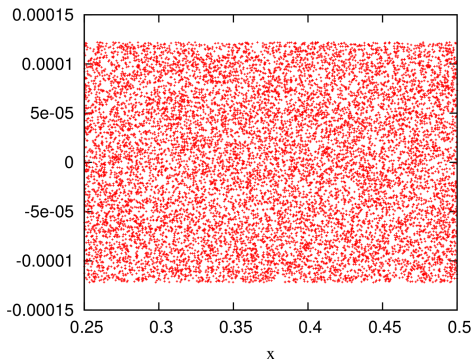
- Stochastic parametrisation schemes use random forcing with specific mean and variability to improve predictability.
- Rounding errors will generate a forcing that is added to the differential equations that is uncorrelated in space and time.

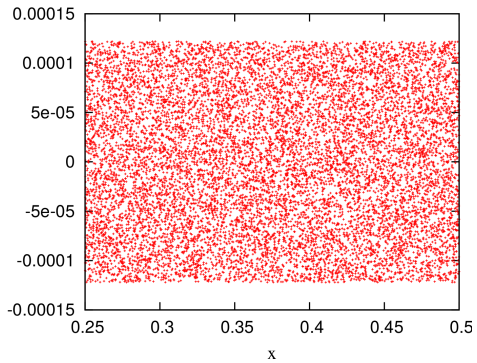




Compare rounding errors with stochastic parametrisation schemes

- Stochastic parametrisation schemes use random forcing with specific mean and variability to improve predictability.
- Rounding errors will generate a forcing that is added to the differential equations that is uncorrelated in space and time.

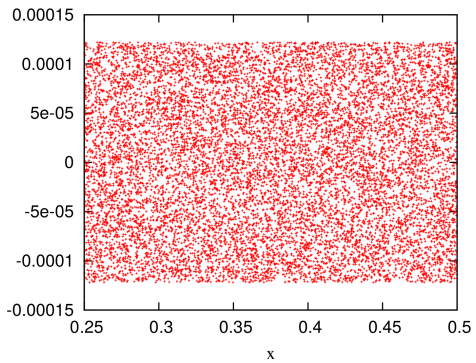


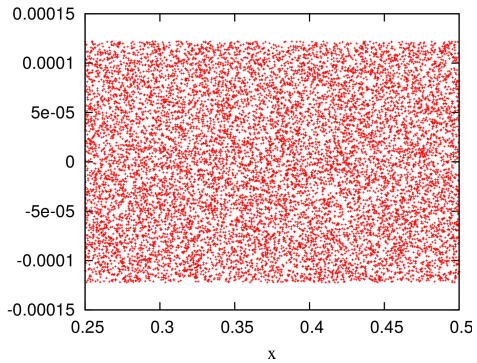


Can we design noise from rounding errors to replace the random forcing of stochastic parametrisation schemes.

Compare rounding errors with stochastic parametrisation schemes

- Stochastic parametrisation schemes use random forcing with specific mean and variability to improve predictability.
- Rounding errors will generate a forcing that is added to the differential equations that is uncorrelated in space and time.

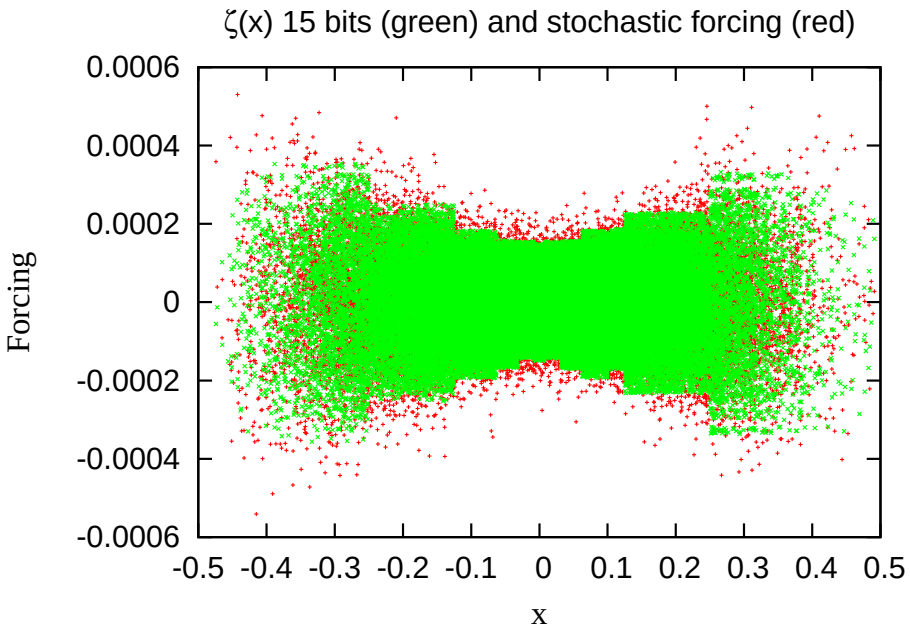


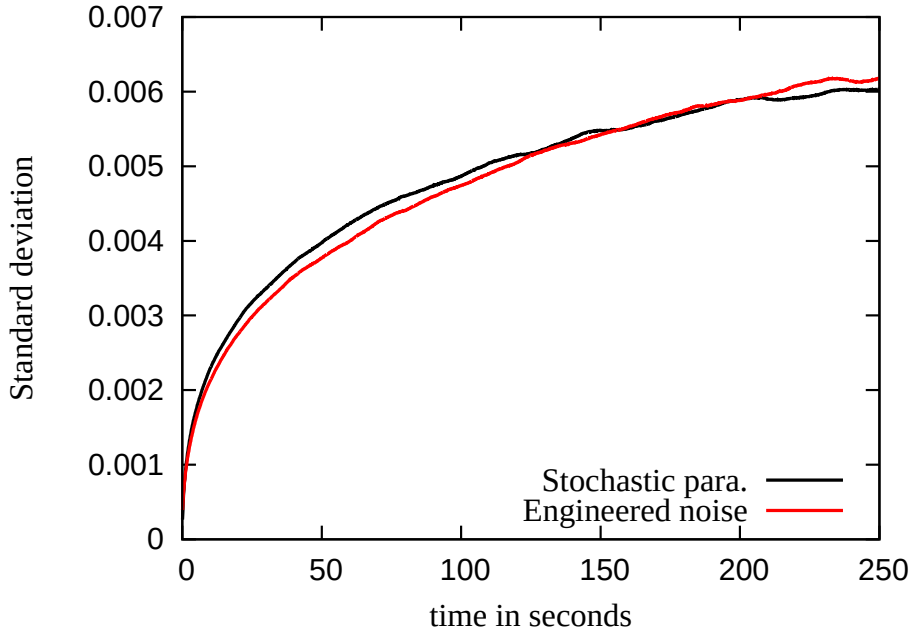


Can we design noise from rounding errors to replace the random forcing of stochastic parametrisation schemes.

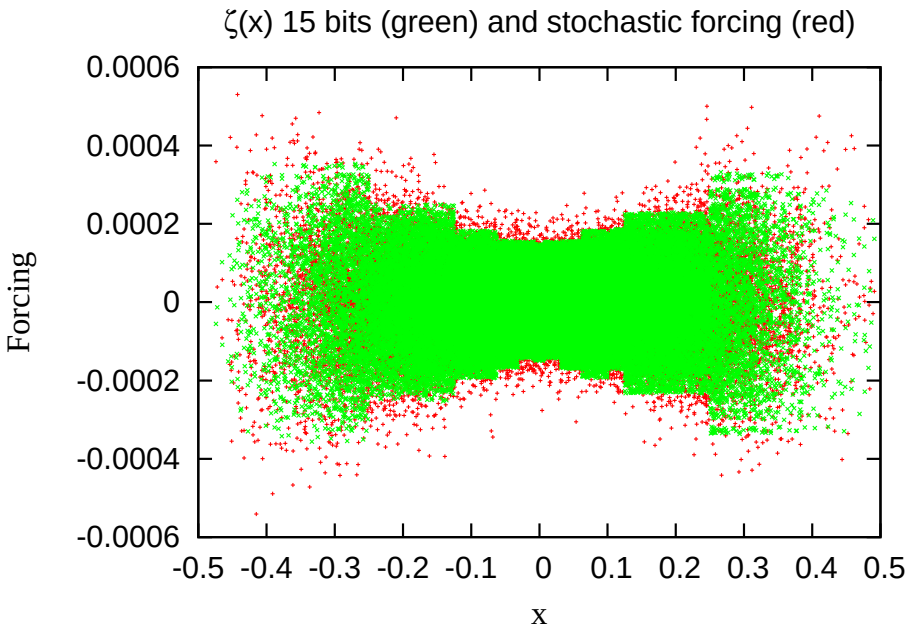
We study a Burgers equation model with stochastic turbulent closure scheme.

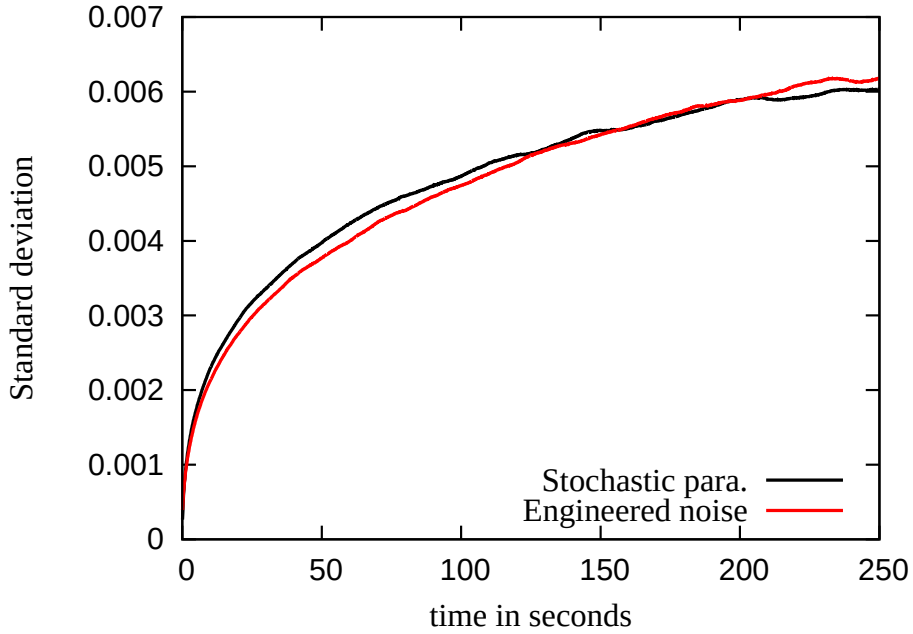
Compare ronding errors with stochastic parametrisation schemes





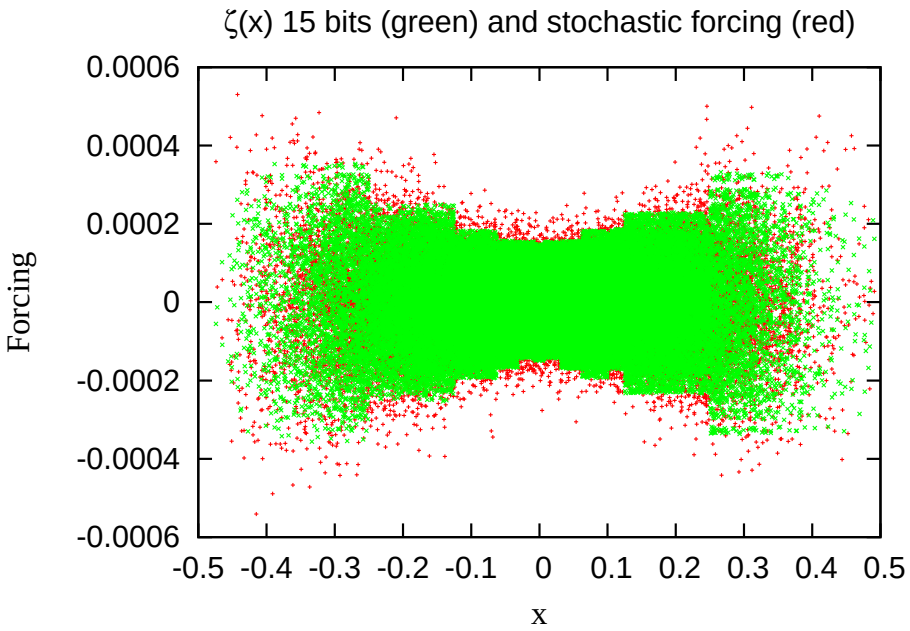
Compare ronding errors with stochastic parametrisation schemes

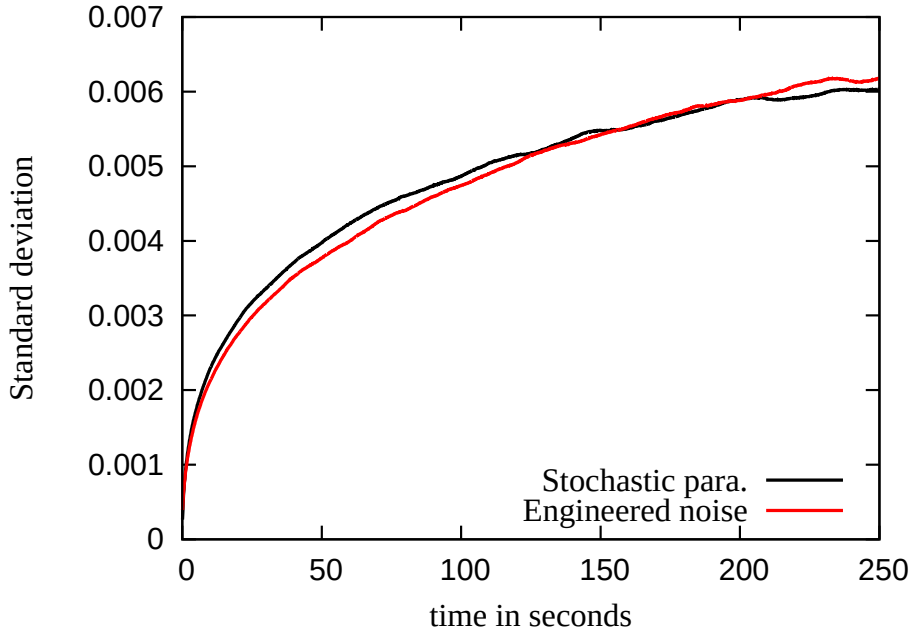




Rounding errors can be hidden by stochastic parametrisation schemes.

Compare ronding errors with stochastic parametrisation schemes

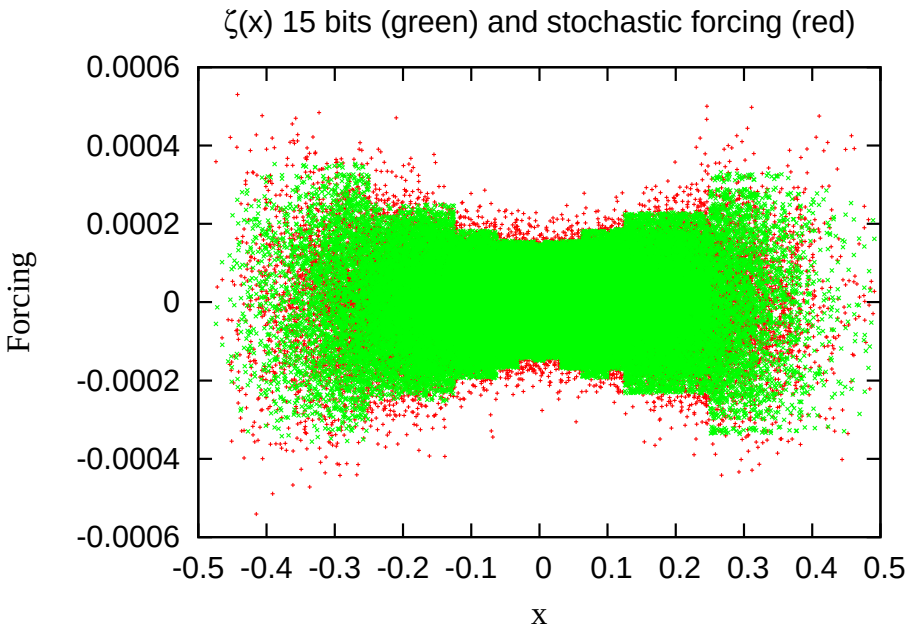


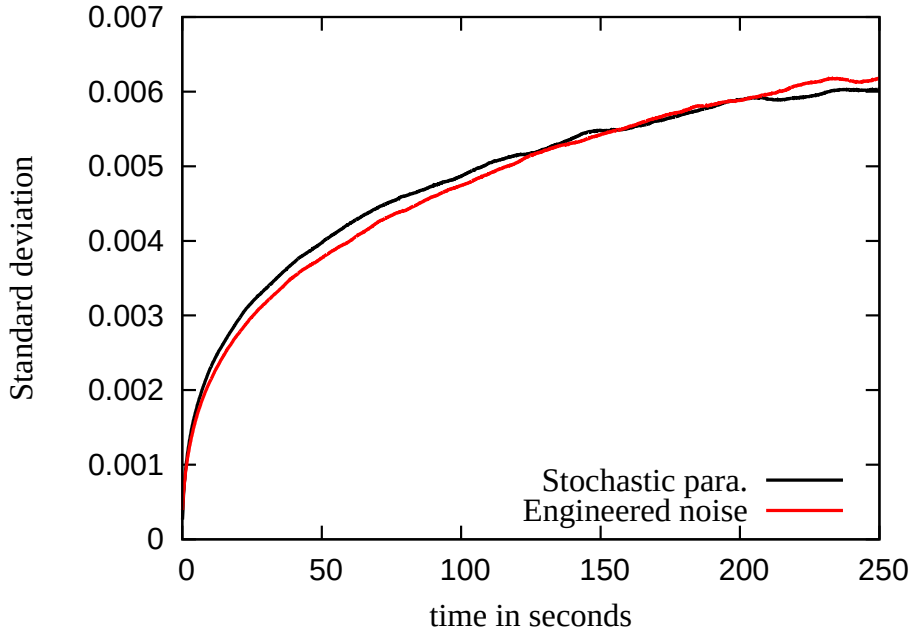


Rounding errors can be hidden by stochastic parametrisation schemes.

Rounding errors can represent sub-grid-scale variability.

Compare ronding errors with stochastic parametrisation schemes





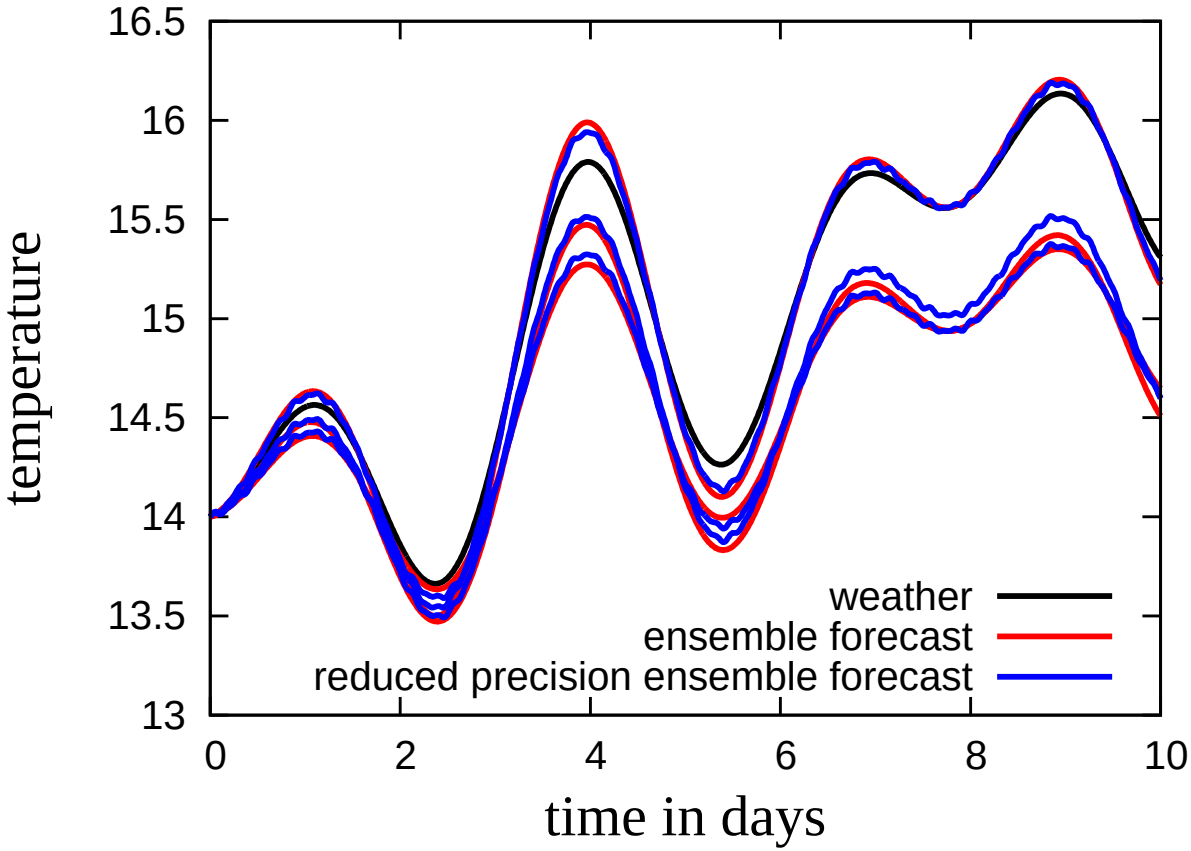
Rounding errors can be hidden by stochastic parametrisation schemes.

Rounding errors can represent sub-grid-scale variability.

This study is extremely idealized.

Düben and Dolaptchiev TCFD 2015

Two research questions



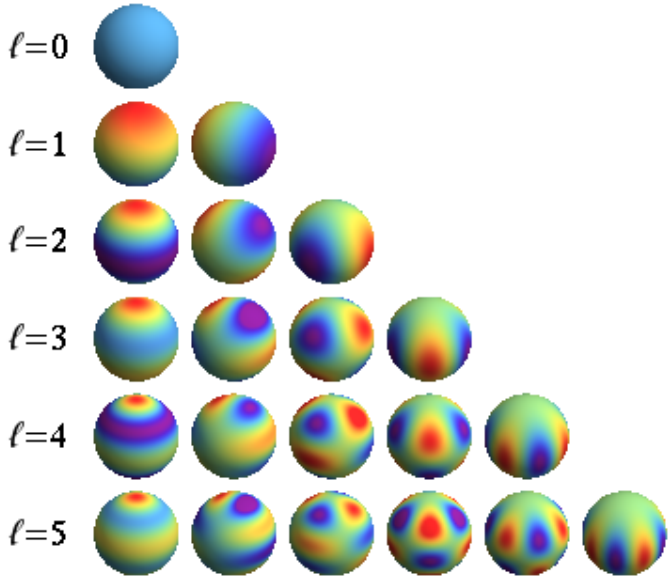
Will our models fail if we reduce precision? - No!

Can we identify the optimal level of precision? - Yes!

One more research questions

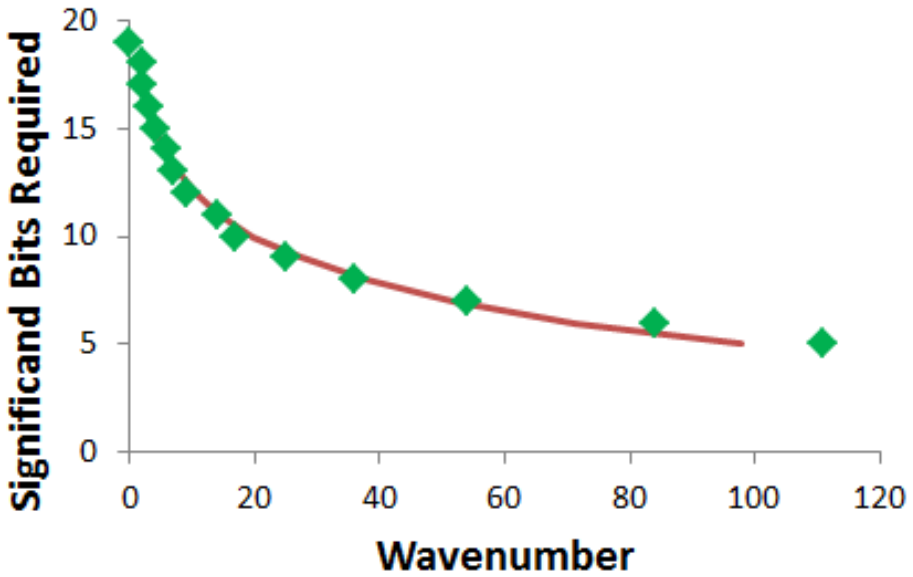
Can a study of numerical precision help to understand model uncertainty and model error?

A scale-selective approach



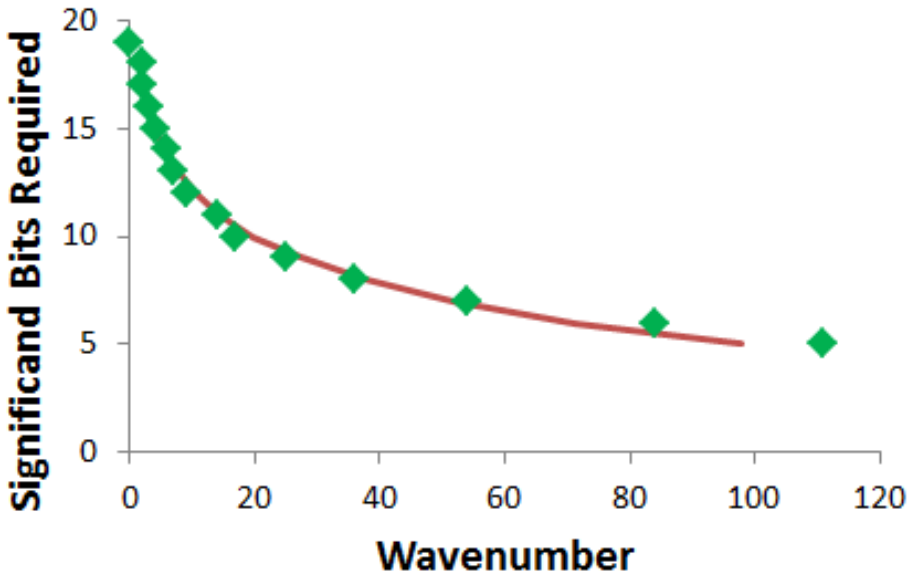
- Spectral models allow to treat different scales at different precision.
- We can push the small scales harder than the large scales.
- This is intuitive due to the high inherent uncertainty in small scale dynamics (parametrisation, viscosity, data-assimilation,...).
- The smallest scales are most expensive.

A scale-selective approach
PhD student Tobias Thornes



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

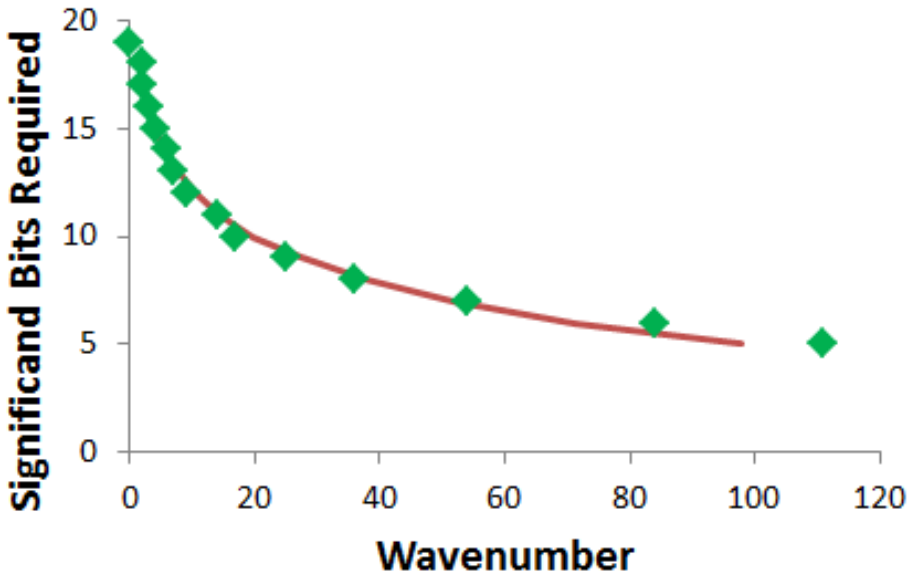
A scale-selective approach
PhD student Tobias Thornes



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

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

A scale-selective approach
PhD student Tobias Thornes



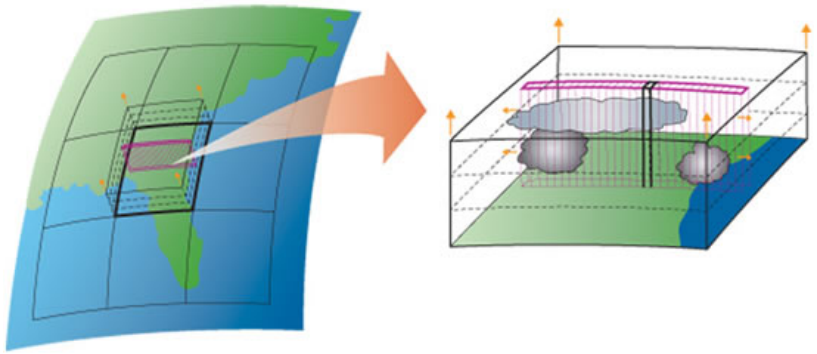
A scale-dependent reduction in precision in the surface quasi-geostropic equations.

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

Scale dependent levels of rounding errors should be used to develop stochastic parametrisation schemes.

Thornes, Düben and Palmer QJRMS 2017

Analyse precision to learn about error and uncertainty



- Superparametrisation is running a two-dimensional cloud resolving model in each grid-cell of a global simulation.
- Superparametrisation improves tropical predictions but it is very expensive.
- We integrate the cloud resolving model using emulated reduced precision.

Figure source: http://www.ucar.edu/communications/quarterly/summer06/cloudcenter.jsp

Analyse precision to learn about error and uncertainty

- We automate the search for reduced precision to find the optimal level of precision for individual parameters and model fields.
- We compare model errors due to reduced precision with ensemble spread.

Analyse precision to learn about error and uncertainty

- We automate the search for reduced precision to find the optimal level of precision for individual parameters and model fields.
- We compare model errors due to reduced precision with ensemble spread.

Parameter/Variable	Precision	Error
specific heat of air	7	0.000%
gravitational acceleration	7	0.025%
gas constant water vapour	8	0.000%
diffusivity water vapour	7	0.209%
dynamic viscosity of air	3	0.022%
sub-grid-scale eddy viscosity	3	6.250%
zonal wind	17	$3.81 \cdot 10^{-4}$%
moist static energy	23	$5.96 \cdot 10^{-6}$%
pressure	22	$1.19 \cdot 10^{-5}$
temperature	23	$5.96 \cdot 10^{-6}$%
water vapour	17	$3.81 \cdot 10^{-4}$%
...		

Analyse precision to learn about error and uncertainty

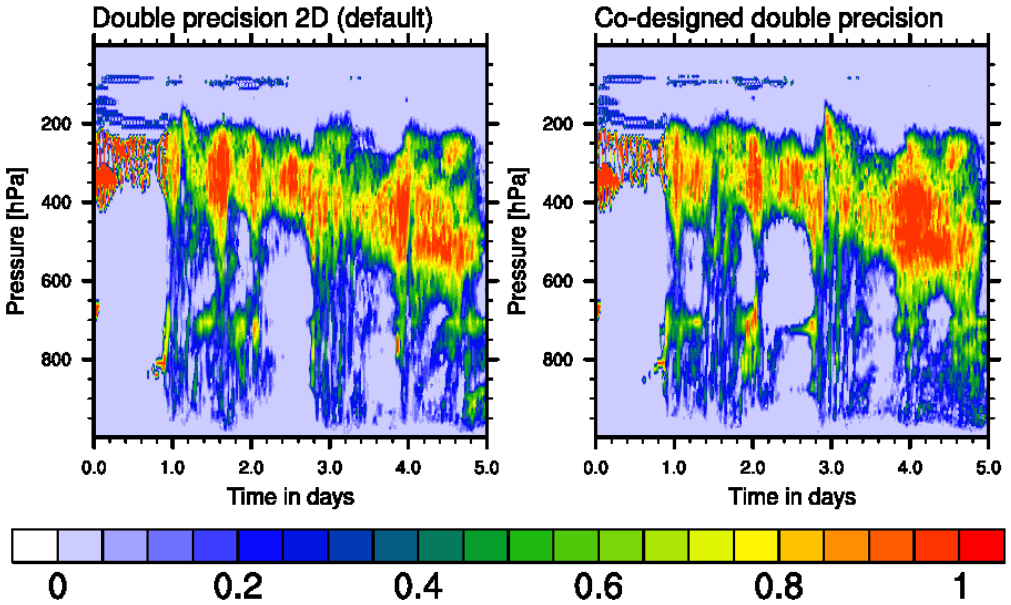
- We automate the search for reduced precision to find the optimal level of precision for individual parameters and model fields.
- We compare model errors due to reduced precision with ensemble spread.

Parameter/Variable	Precision	Error
specific heat of air	7	0.000%
gravitational acceleration	7	0.025%
gas constant water vapour	8	0.000%
diffusivity water vapour	7	0.209%
dynamic viscosity of air	3	0.022%
sub-grid-scale eddy viscosity	3	6.250%
zonal wind	17	$3.81 \cdot 10^{-4}$%
moist static energy	23	$5.96 \cdot 10^{-6}$%
pressure	22	$1.19 \cdot 10^{-5}$
temperature	23	$5.96 \cdot 10^{-6}$%
water vapour	17	$3.81 \cdot 10^{-4}$%
...		

We should use results of the precision analysis to adjust “global” stochastic parametrisation schemes.

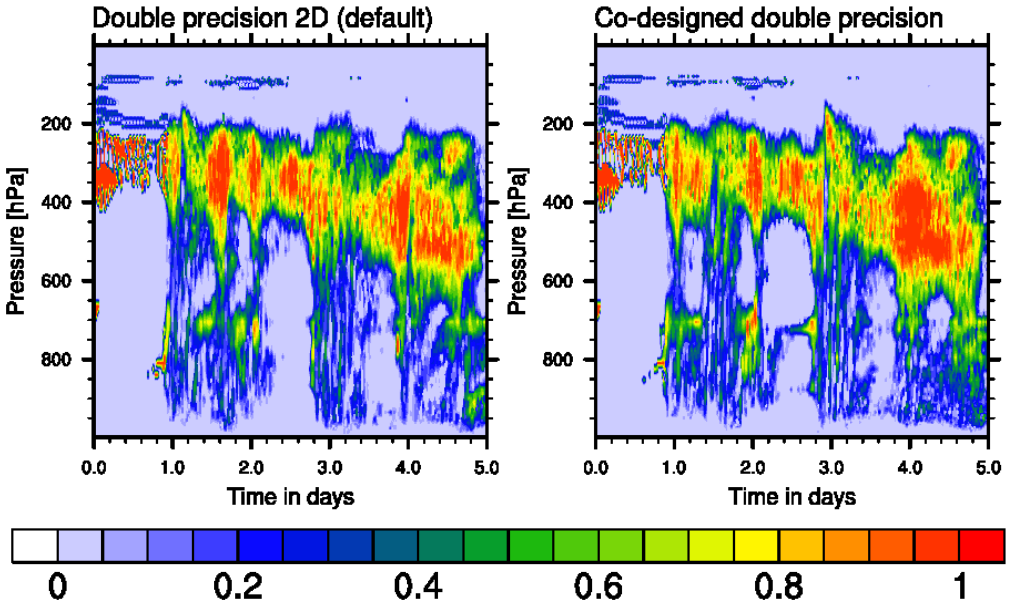
Düben, Subramanian, Dawson and Palmer JAMES 2017

Analyse precision to learn about error and uncertainty



- We find that precision can be reduced significantly in the turbulent kinetic energy scheme and for the high orders of the water vapour saturation curve.
- We remove those parts from the model.
- The new model setup is approximately 12% faster.

Analyse precision to learn about error and uncertainty



- We find that precision can be reduced significantly in the turbulent kinetic energy scheme and for the high orders of the water vapour saturation curve.
- We remove those parts from the model.
- The new model setup is approximately 12% faster.

A precision analysis can help to adjust model complexity.

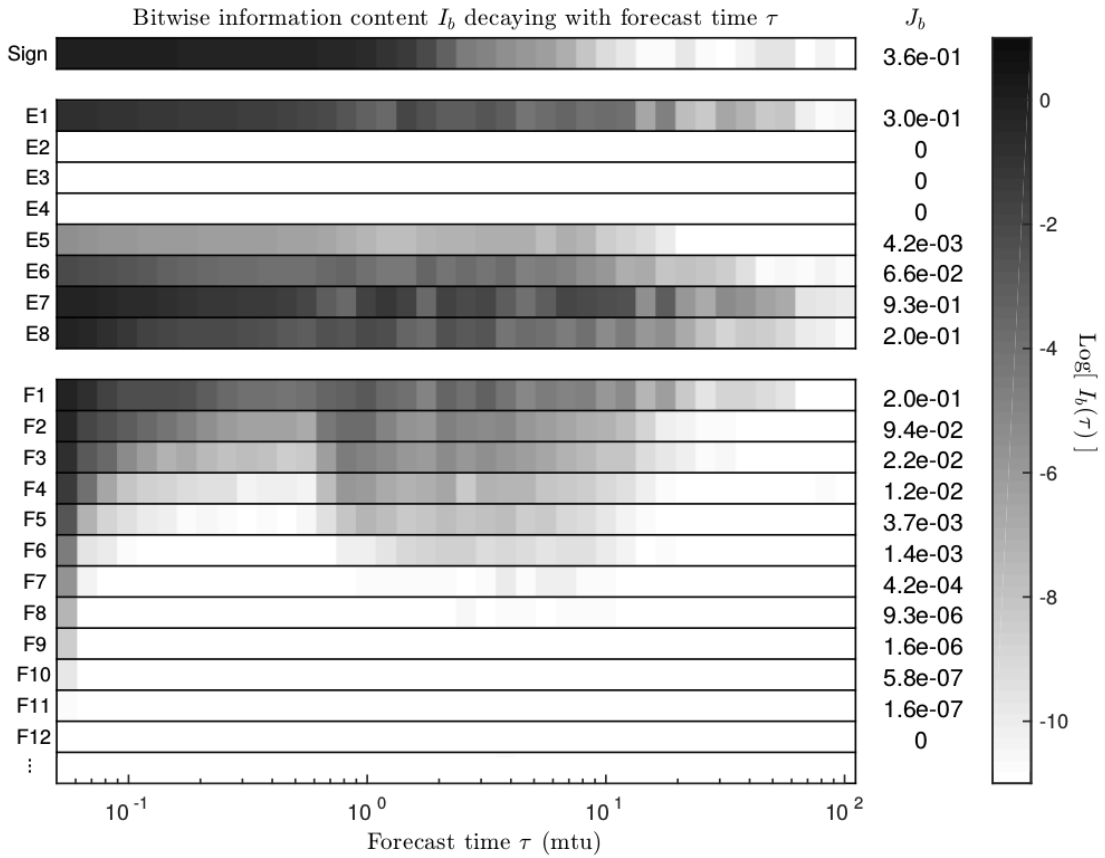
Düben, Subramanian, Dawson and Palmer JAMES 2017

One more research questions

Can a study of numerical precision help to understand model uncertainty and model error?

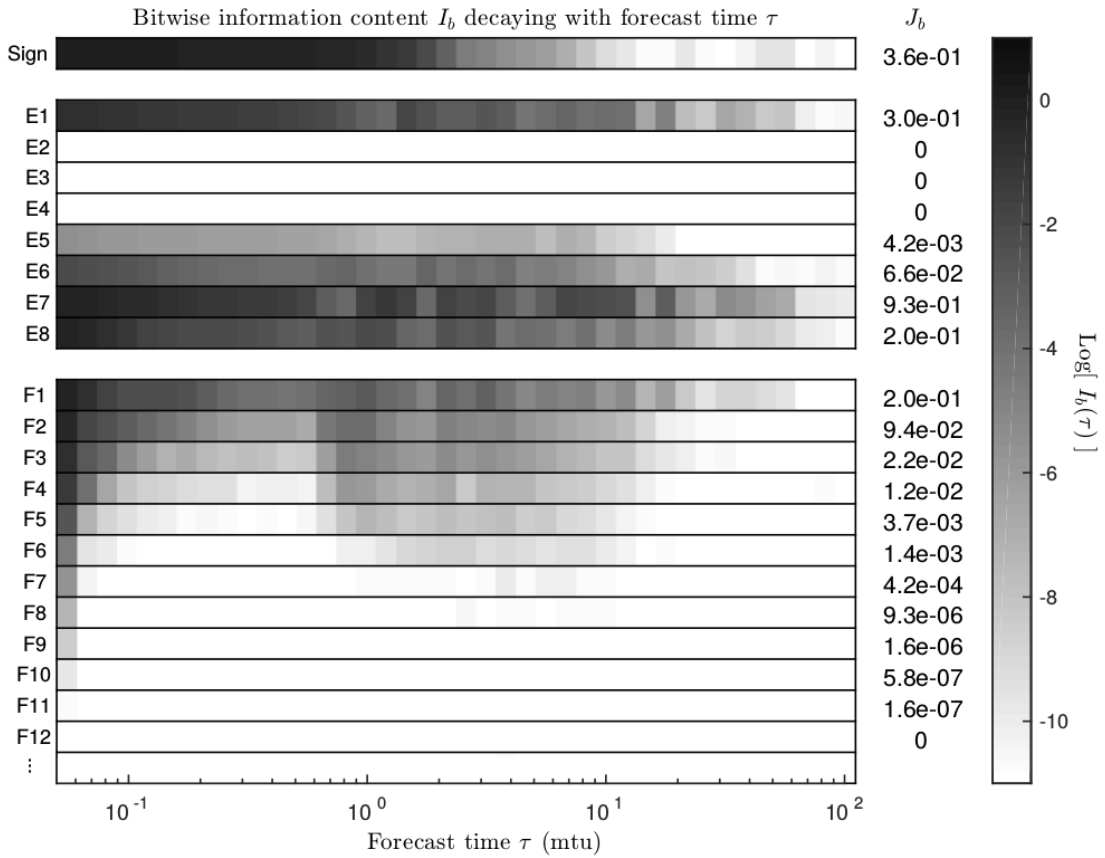
Yes!

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

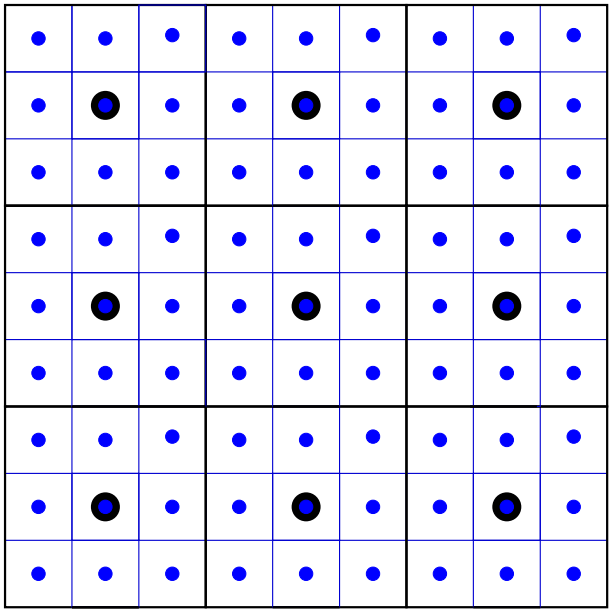


Information content of bits for a Lorenz'63 model using a single long term integration and Shannon information theory.

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

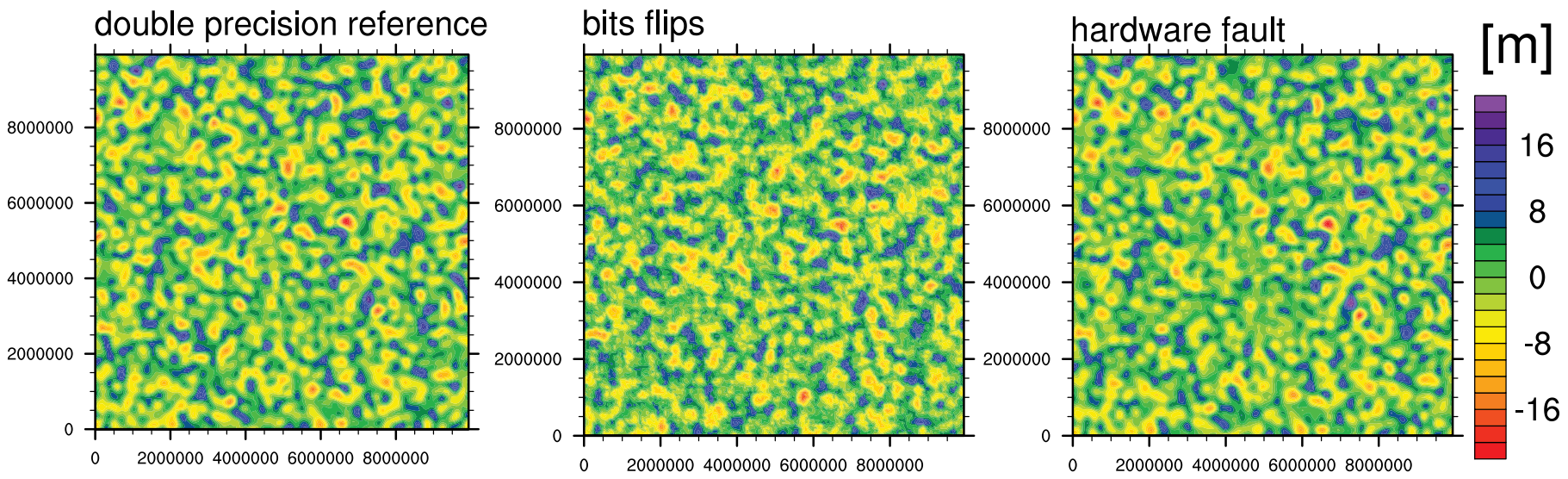
Jeffress, Düben and Palmer in prep. for Proc. R. Soc. A

Shallow water model with hardware faults



- We introduce a coarse backup grid to save prognostic fields.
- We test whether the fields on the backup grids are physically meaningful and restore erroneous values on the model grid, using the backup grid.
- We emulate soft errors in floating point operations and the loss of information in large areas of the model domain.
- The backup system generates 13% overheads.

Shallow water model with hardware faults



- We introduce a coarse backup grid to save prognostic fields.
- We test whether the fields on the backup grids are physically meaningful and restore erroneous values on the model grid, using the backup grid.
- We emulate soft errors in floating point operations and the loss of information in large areas of the model domain.
- The backup system generates 13% overheads.

How to approach full-blown GCMs?
Emulation of reduced precision

Method:
We define a new reduced-precision type that behaves like a floating point number, but reduces the precision when it is operated on, this allows the emulation of reduced precision and specific setups of inexact hardware in large models (maybe IFS?) with no need for extensive changes of model code.

Example:
Emulated 5 bit significand with reduced precision "+"

Standard Fortran:

```
REAL :: a,b,c
a = 1.442221
b = 2.136601
c = a+b
```

→ *c=3.578822*

Reduced precision declarations:

```
TYPE(reduced_precision) :: a,b,c
a = 1.442221
b = 2.136601
c = a+b
```

→ *c=3.562500*

Conclusions

Scientific challenges to improve forecasts:

- The free lunch is over in high performance computing.
- We fail to provide a satisfying representation of model uncertainty in weather and climate models.

Results suggest that...

- a reduction in precision is promising huge savings.
- savings can be reinvested to allow higher resolution/complexity or more ensemble members to improve forecasts.
- our understanding of model error and model uncertainty helps to adjust precision.
- precision should be reduced with spatial scale and forecast lead time.
- a precision analysis helps to understand model uncertainty and to adjust stochastic parametrisation schemes.