Reduced Precision Computing and Machine Learning for Earth System Modelling

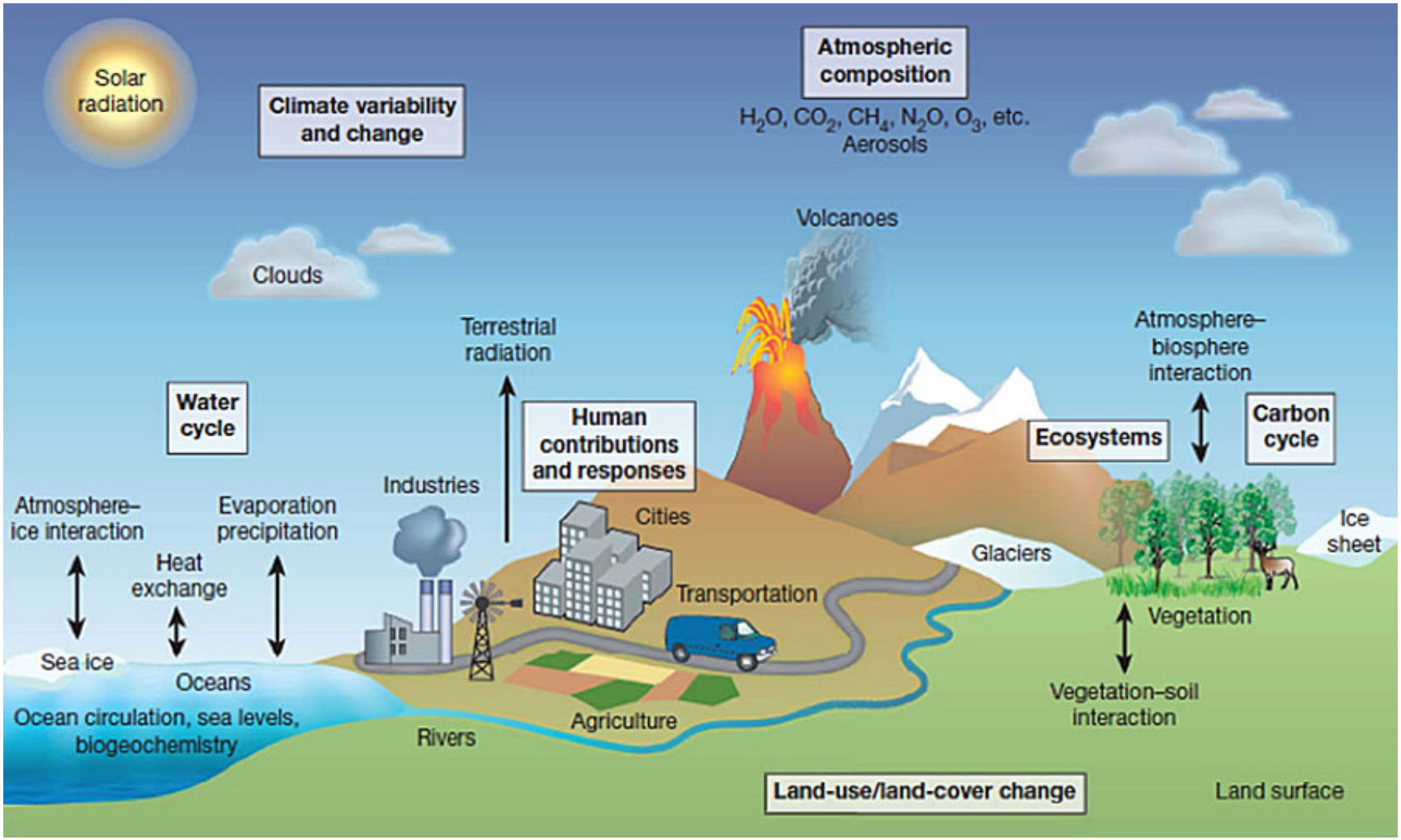
Peter Düben

University Research Fellow of the Royal Society

European Centre for Medium-Range Weather Forecasts (ECMWF)

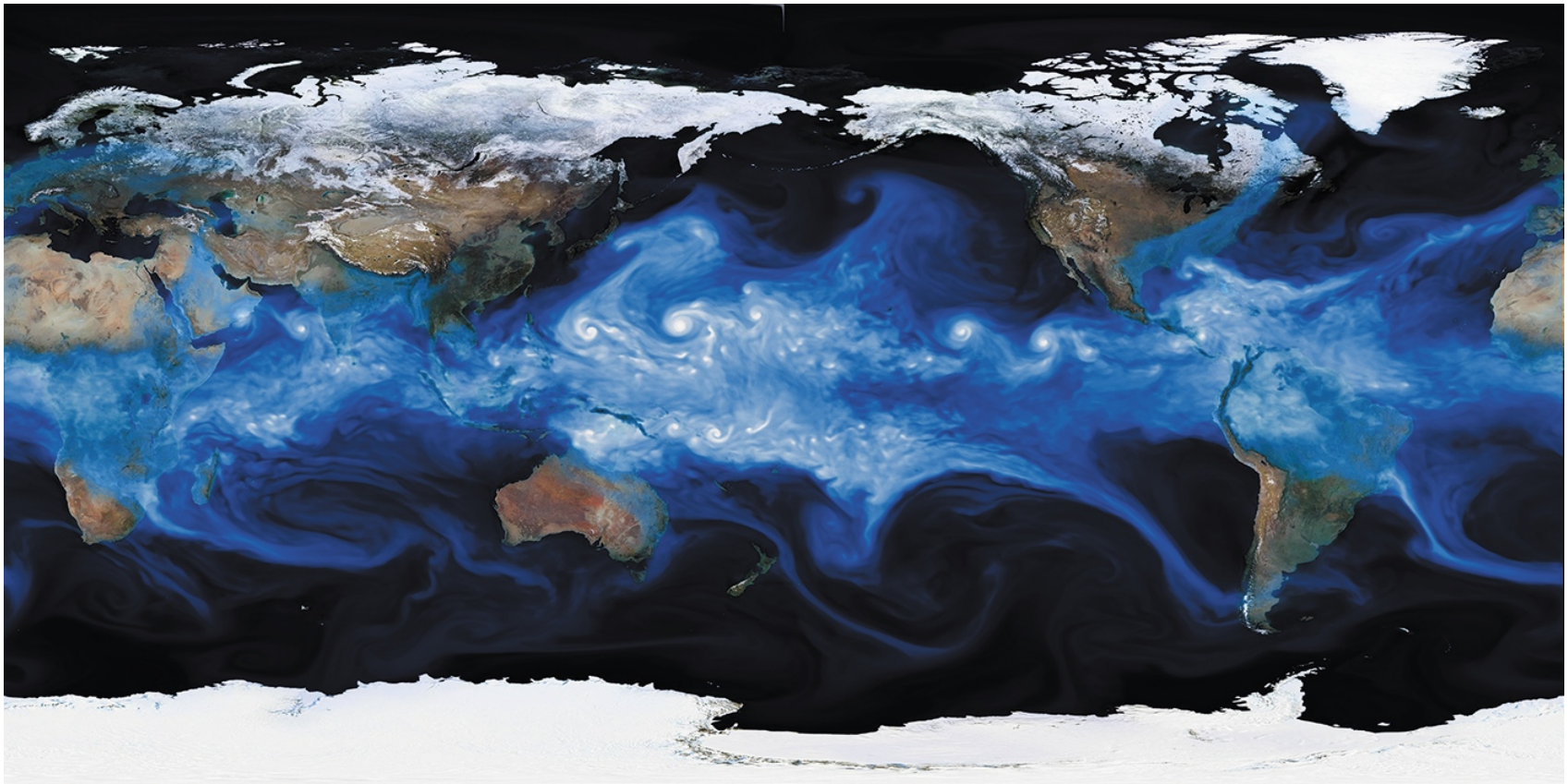
Predicting weather and climate: Why is it so hard?

Predicting weather and climate: Why is it so hard?



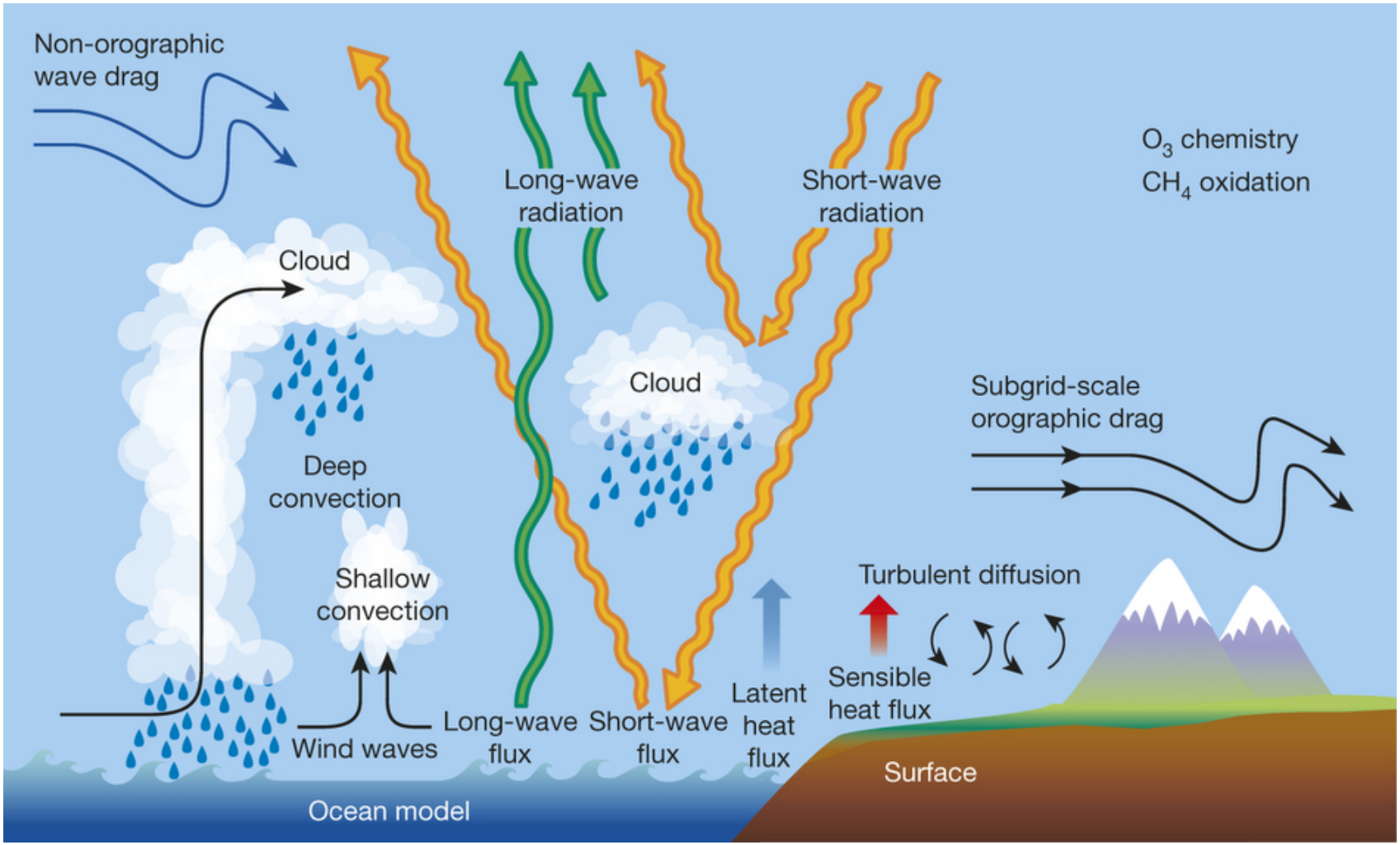
www.gfdl.noaa.gov

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?

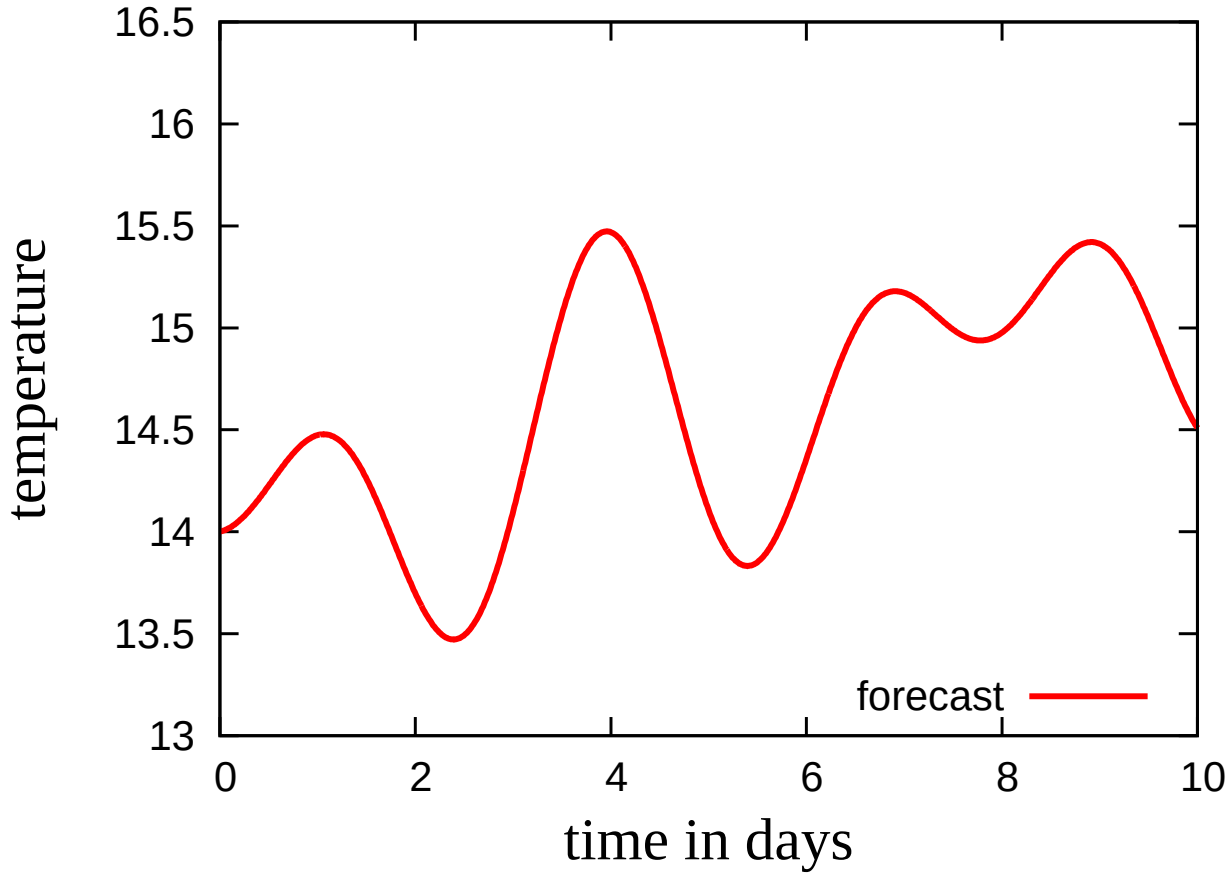
National Geographic Creative

Predicting weather and climate: Why is it so hard?

National Geographic Creative

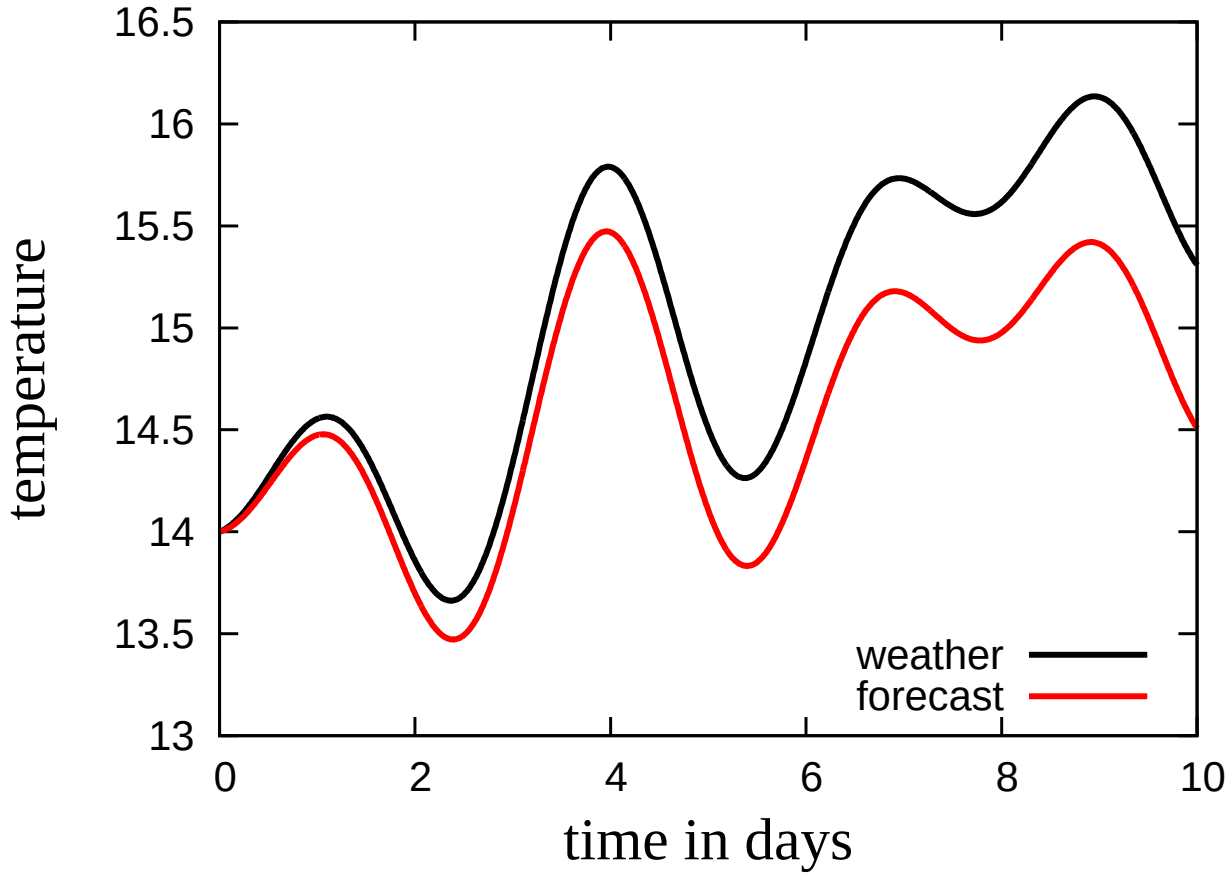
The Earth System is complex, huge and chaotic and we do not have sufficient resolution to resolve all important processes.

How do we treat uncertainties in weather forecasts?



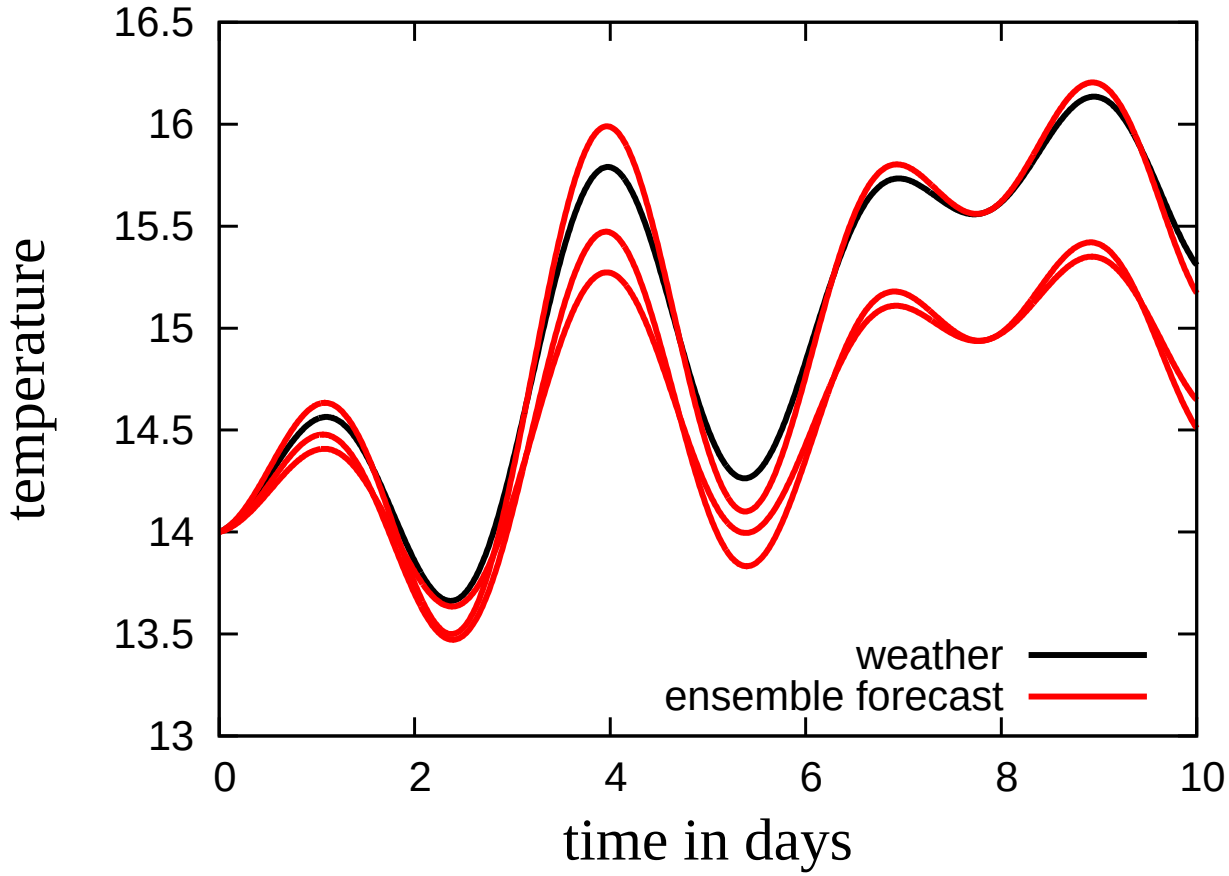
How do we know if we are wrong?

How do we treat uncertainties in weather forecasts?



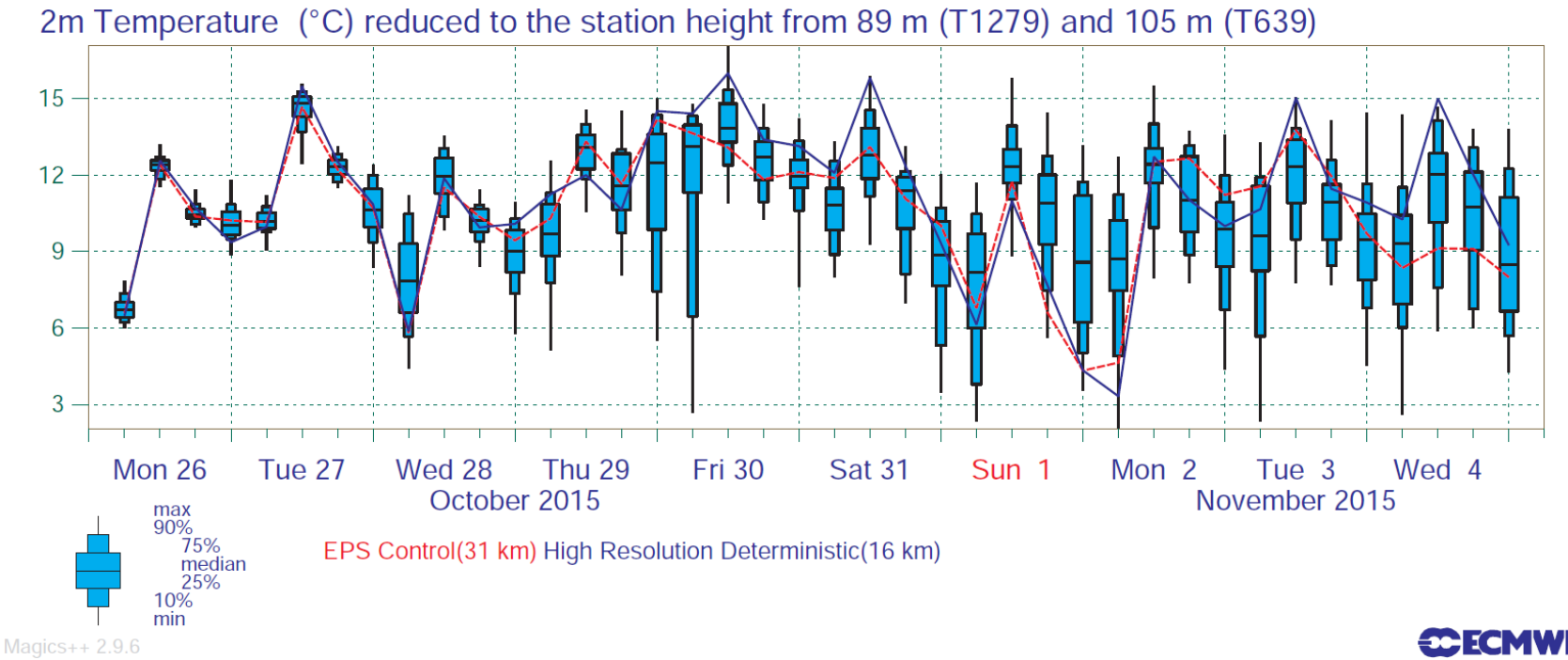
How do we know if we are wrong?

How do we treat uncertainties in weather forecasts?

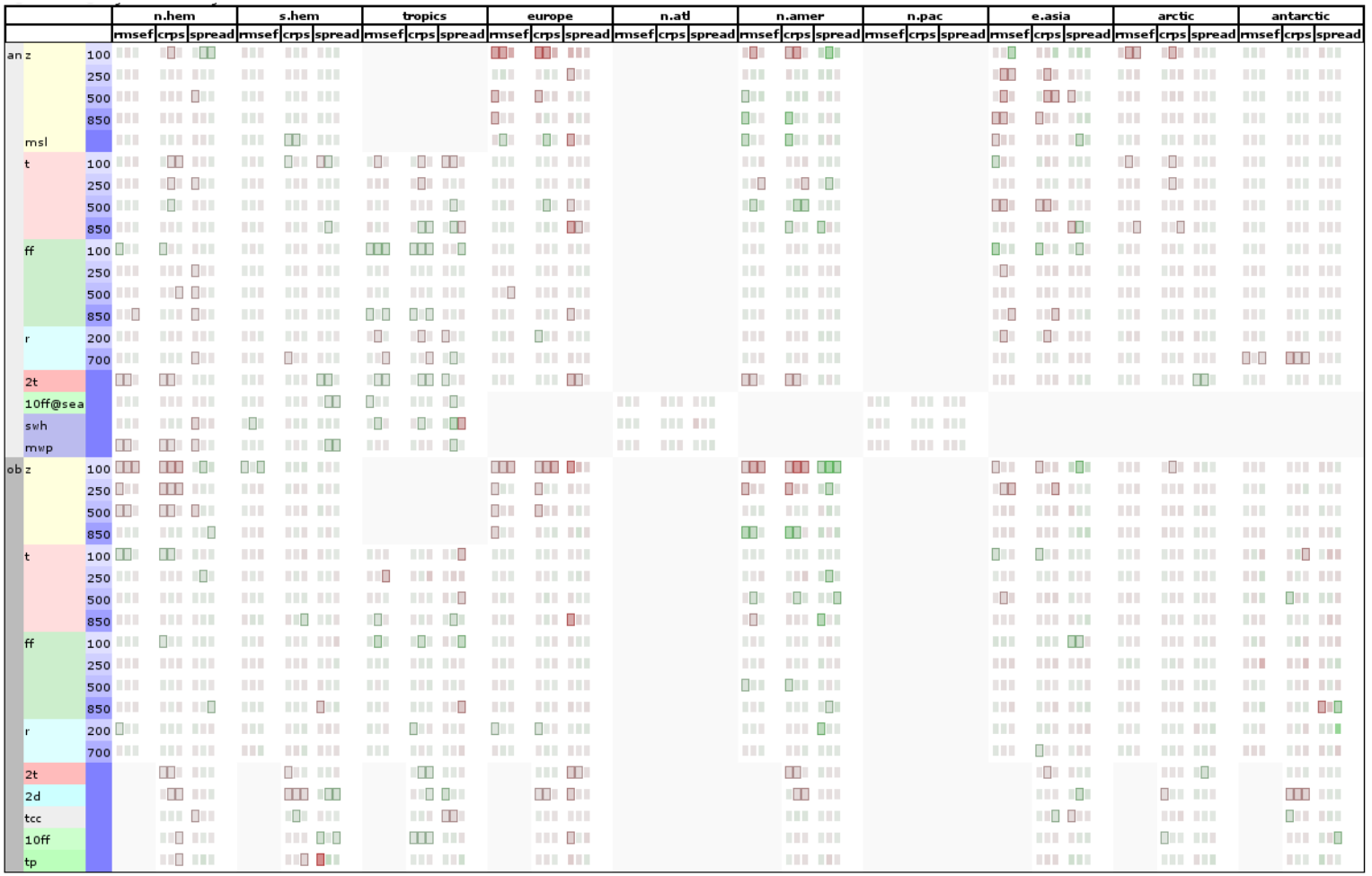


The ensemble spread holds information about forecast uncertainty.

How do we treat uncertainties in weather forecasts?

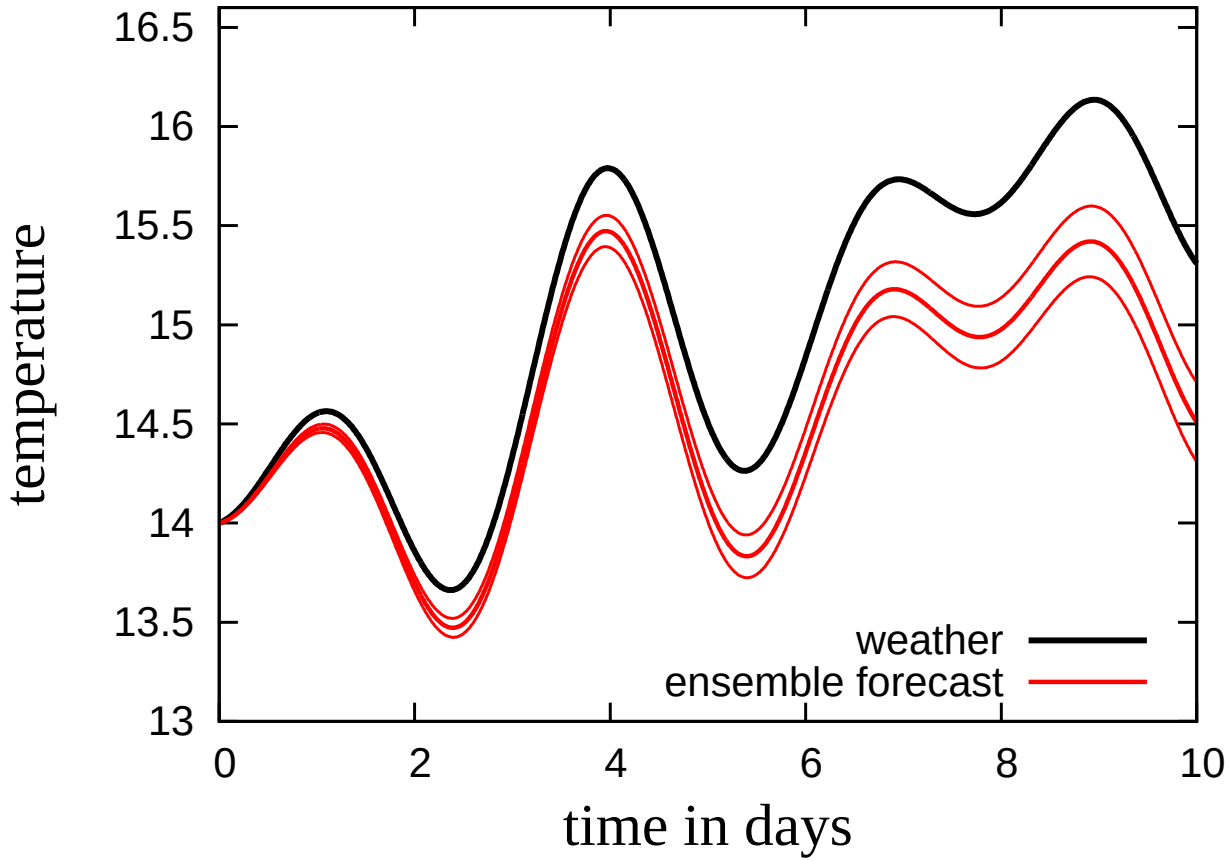


To improve a multi-dimensional, non-linear system...



You may need to run 100 years of a coupled climate model to identify a response to a forcing...

How do we treat uncertainties in weather forecasts?

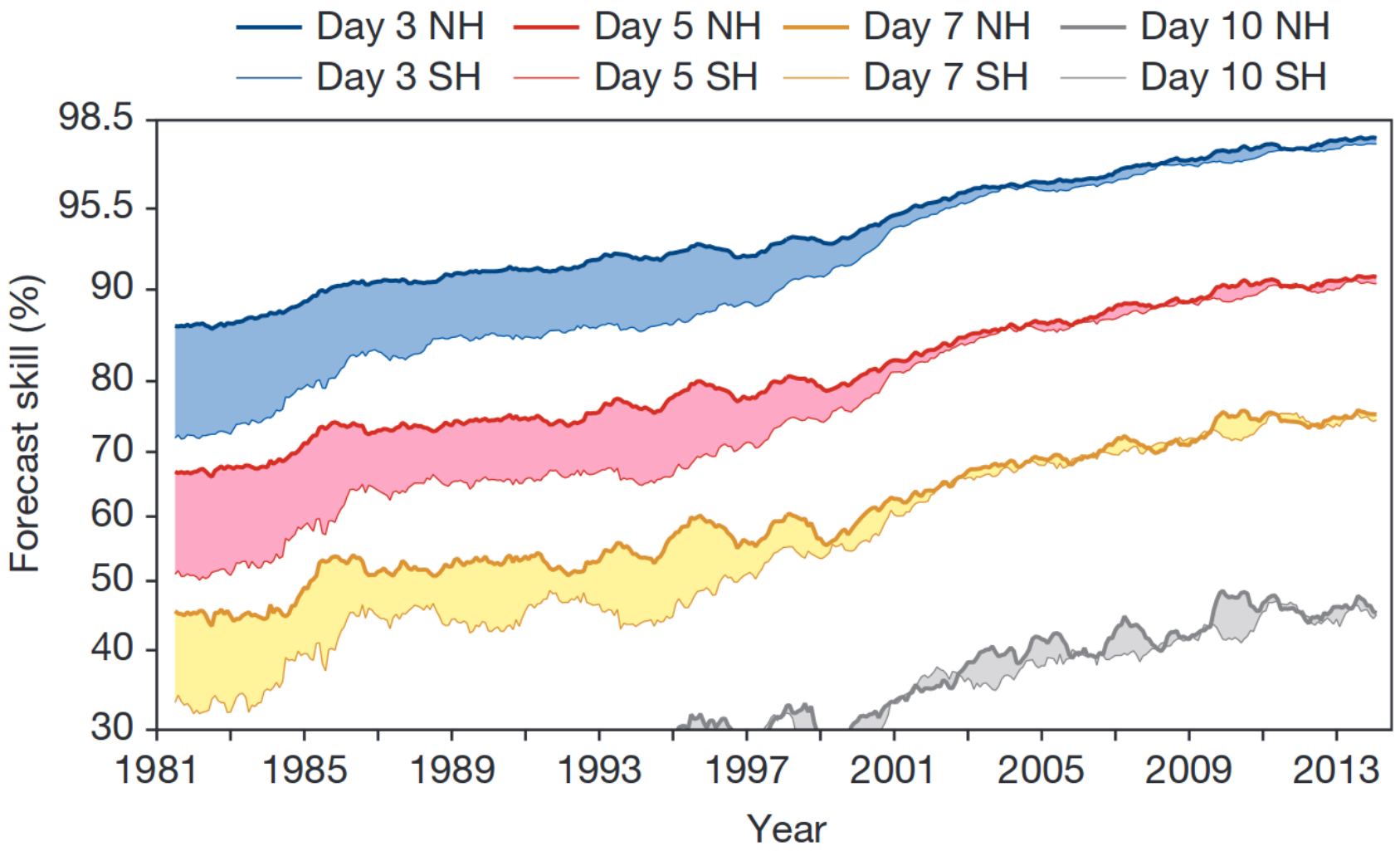


Ensemble forecasts can go wrong.

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

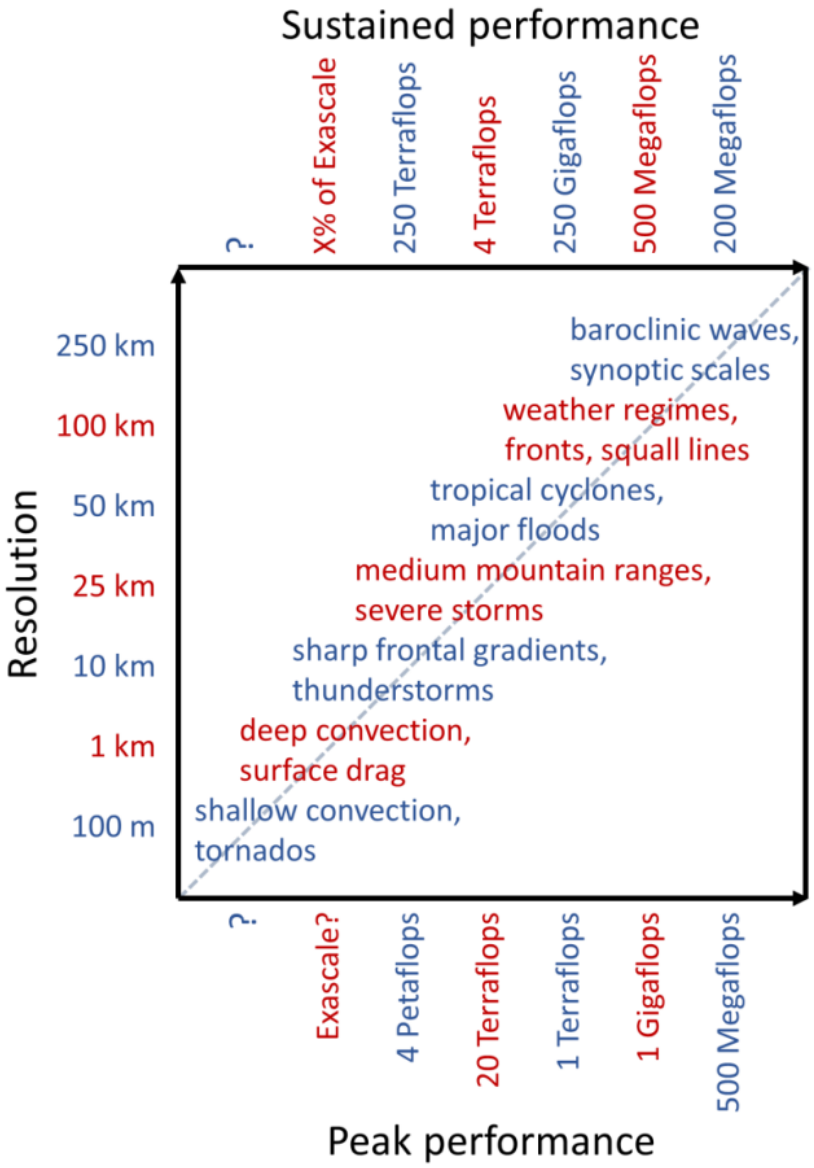
These schemes are typically “local” and lack physical justification.

Forecast skill is still improving



Higher resolution in weather models → improved forecast skill.

Weather and climate models are HPC applications

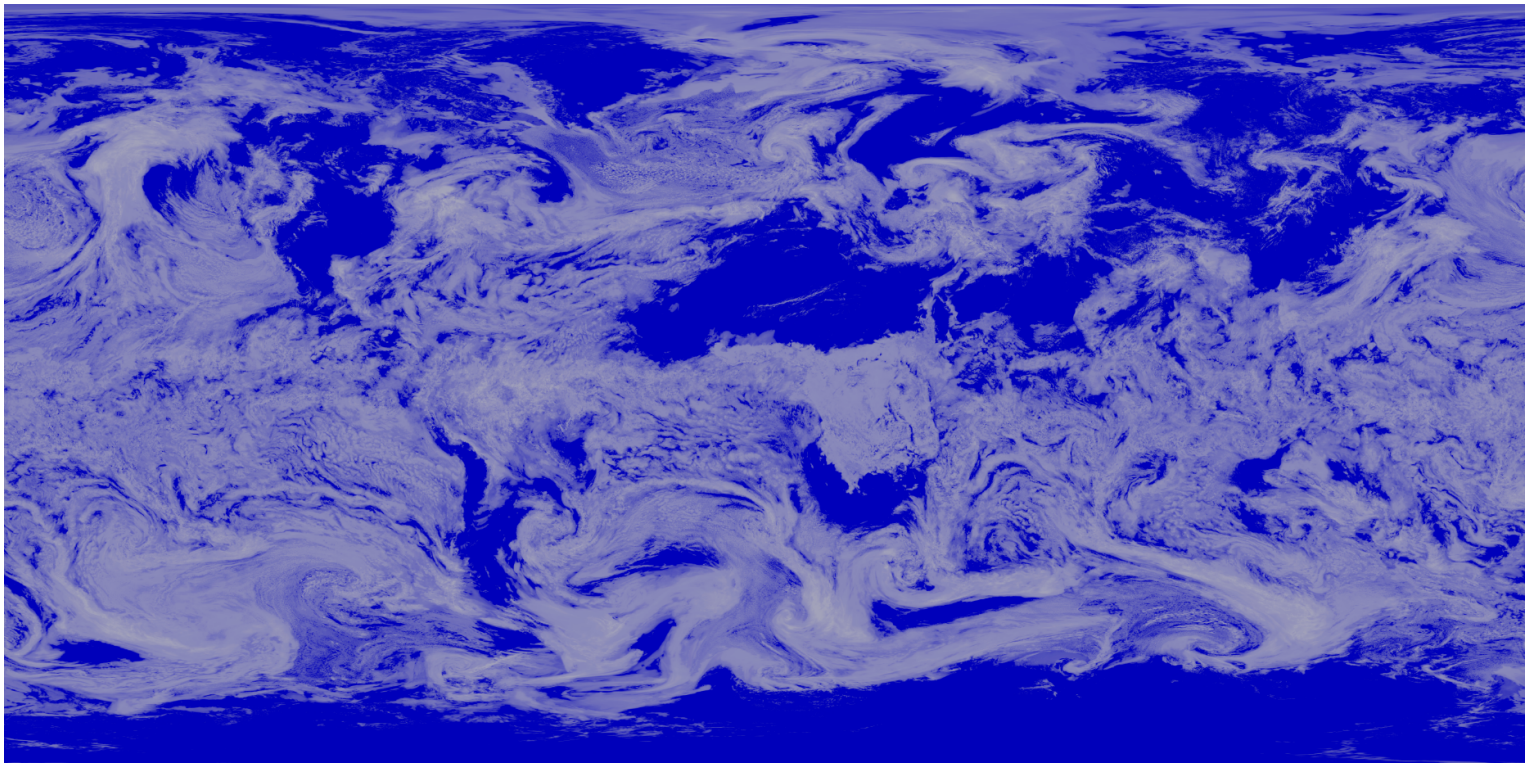


Neumann et al., Phil. Trans. A, 2019

- More resolution
  $\rightarrow$ more processes resolved.
- Ratio sustained/peak is going down.
- 1km resolution allows the explicit representation of deep convection and the generation of gravity waves in the atmosphere and meso-scale eddies in the ocean.

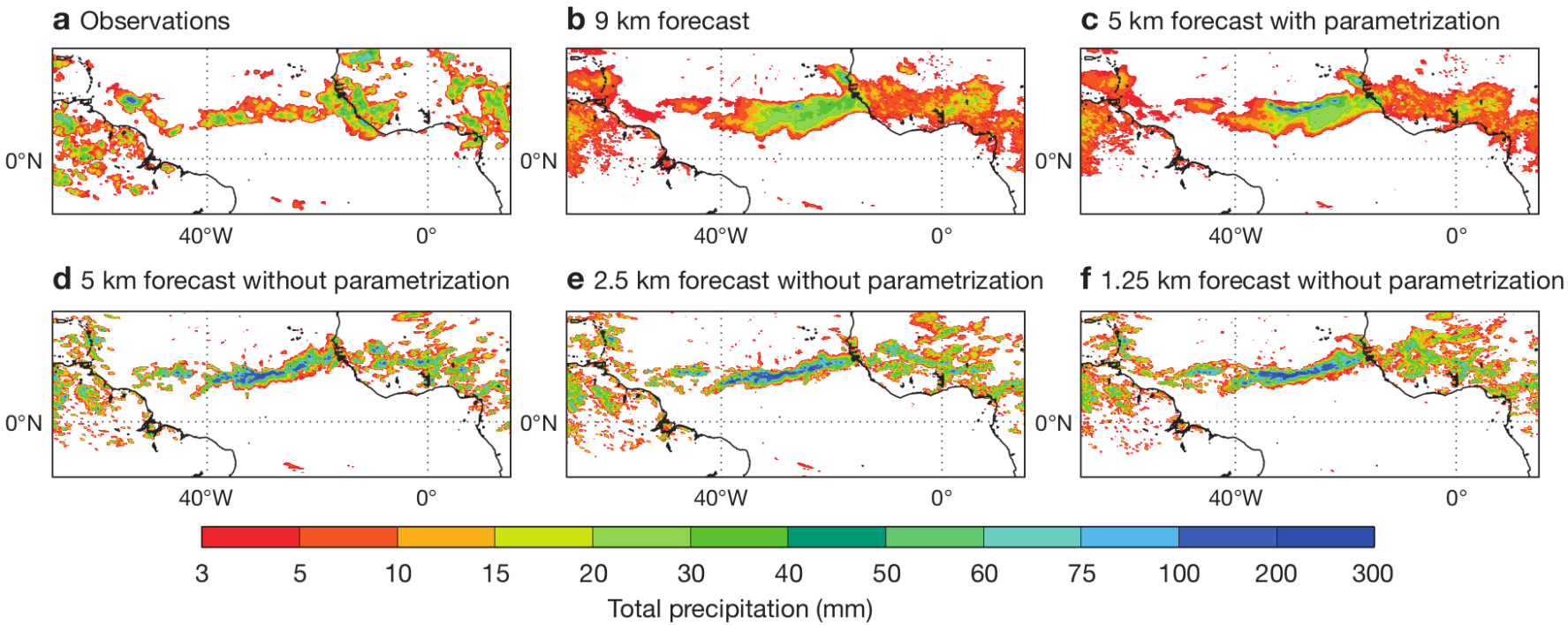
Forecasts with billions of degrees-of-freedom

Clouds in a global weather simulation at 1 km resolution
(Figure courtesy of Nils Wedi)



Global simulations show a breath-taking level of complexity and can represent many details of the Earth System.

An example for the impact of high resolution



- Total precipitation accumulated for 24 hours.
- Explicit representation of deep convection
  $\rightarrow$ more realistic but too strong.
- The simulations were performed as part of the ESiWACE H2020 Centre of Excellence.

Dueben et al. ECMWF Newsletter 2018

Challenges for future High Performance Computing

Technical:

- Individual processors will not be faster.
  - Parallelisation ($> 10^6$ parallel processing units).
  - Power consumption will be a big problem.
  - Hardware faults may jeopardize large simulations.
- Hardware will be more heterogeneous.
  - CPUs, GPUs, FPGAs, ASICs.
  - Different hardware will require different code changes.
  - We do not know what hardware we will be using in 10 years.
  - We need Domain Specific Languages to port models.
- Machine learning has a strong impact on hardware development. High floprate at low precision (16 bits and lower).
- I/O will (need to) become a focus.

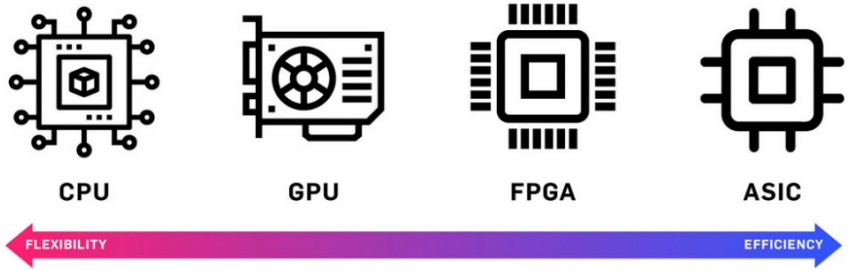
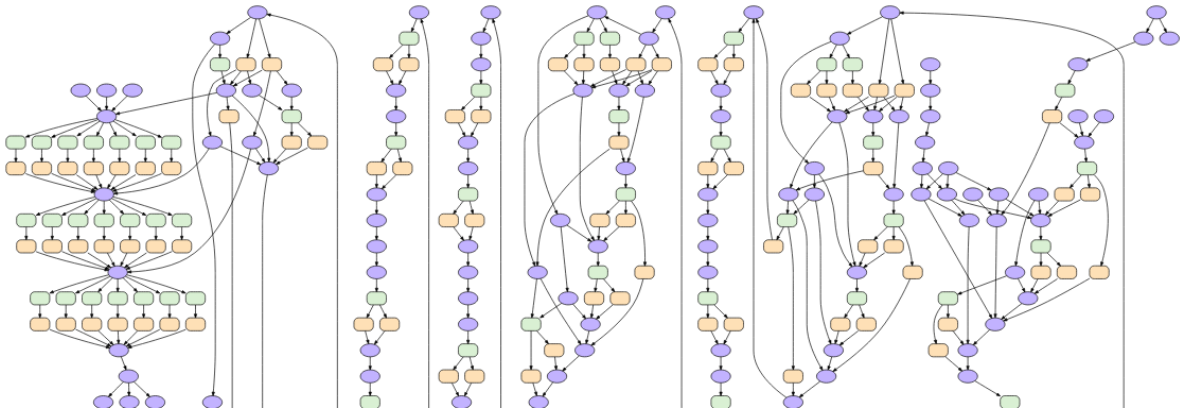


Figure copied from https://venturebeat.com

Challenges for future High Performance Computing

Scientific:

- Build algorithms that are efficient and accurate, and scale.
  Explicit vs. implicit; spectral vs. gridpoint; ...
- Build performance models that are required for co-design.
- Optimise data flow and use in the models (NP-hard problem).
- Tune and evaluate high-resolution simulations.
  Response to forcing of a non-linear system.
- Adjust accuracy to model uncertainty.



Fuhrer et al. GMD 2018

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

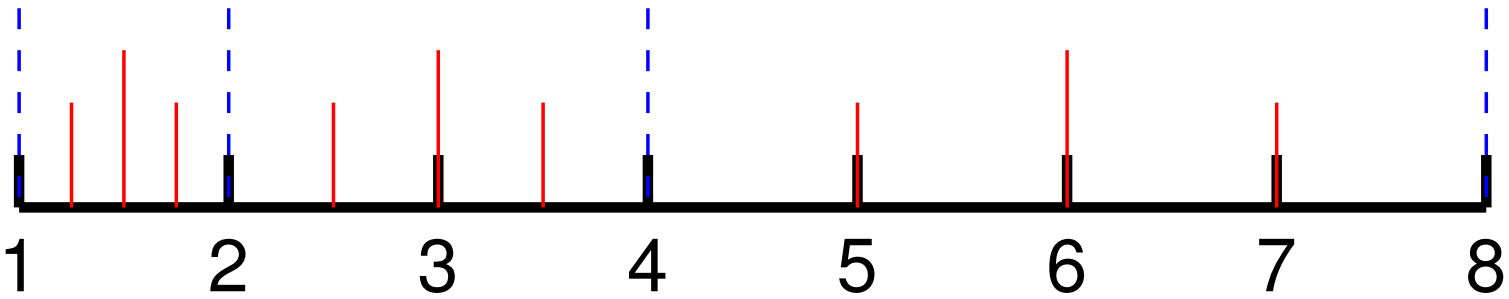
A short introduction to bit representation

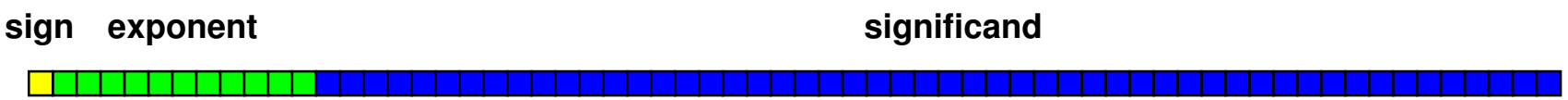
- The computer represents an integer number as a string of 32 bits. Each bit represents a power of two:

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

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

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

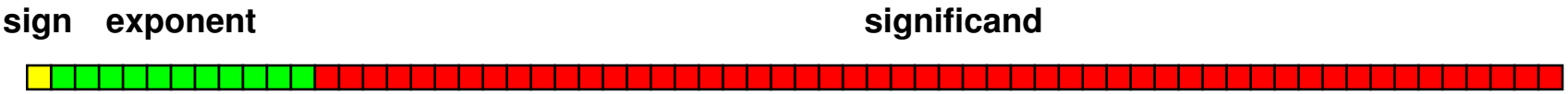




Approaches to inexact floating point units

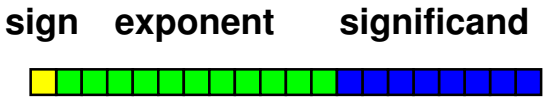
Stochastic processor

- 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.



Pruning

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

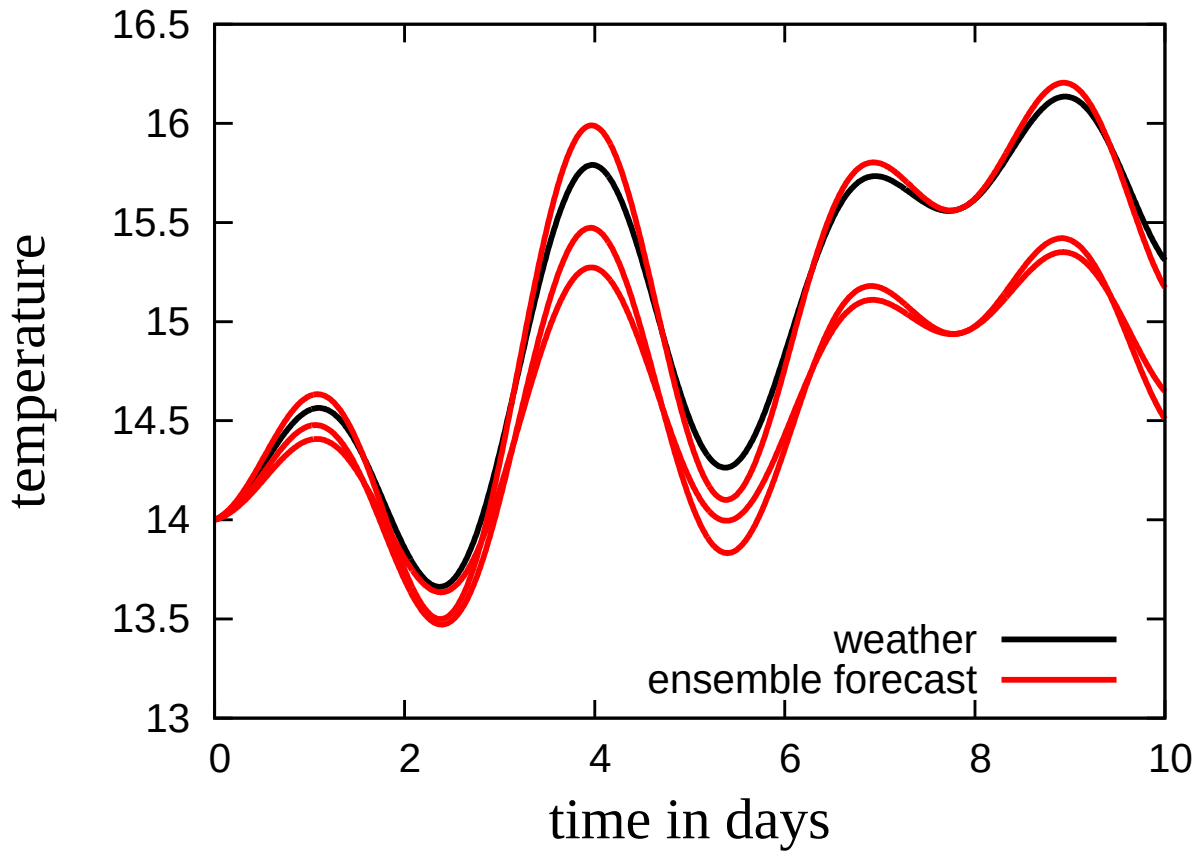


Field Programmable Gate Array (FPGA)

- 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.

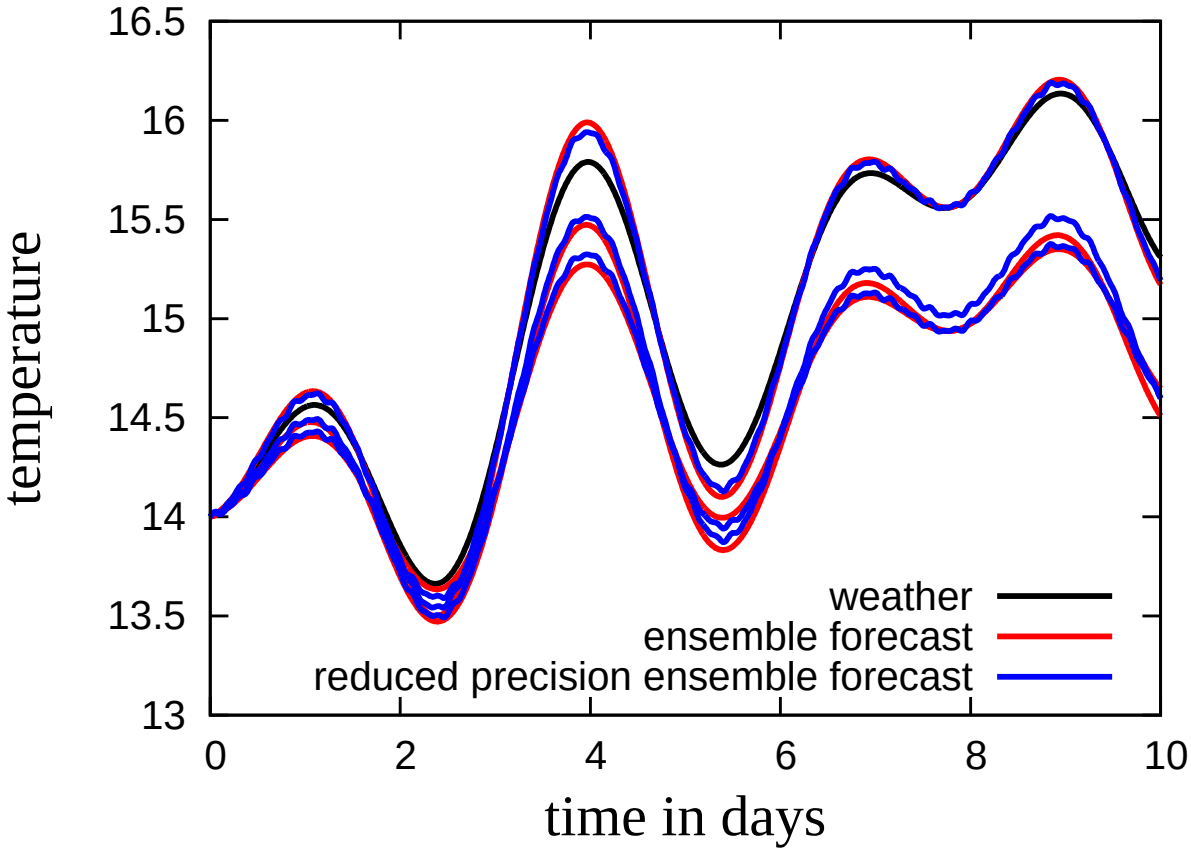
Two research questions



Will our models fail if we reduce precision?

Can we identify the optimal level of precision?

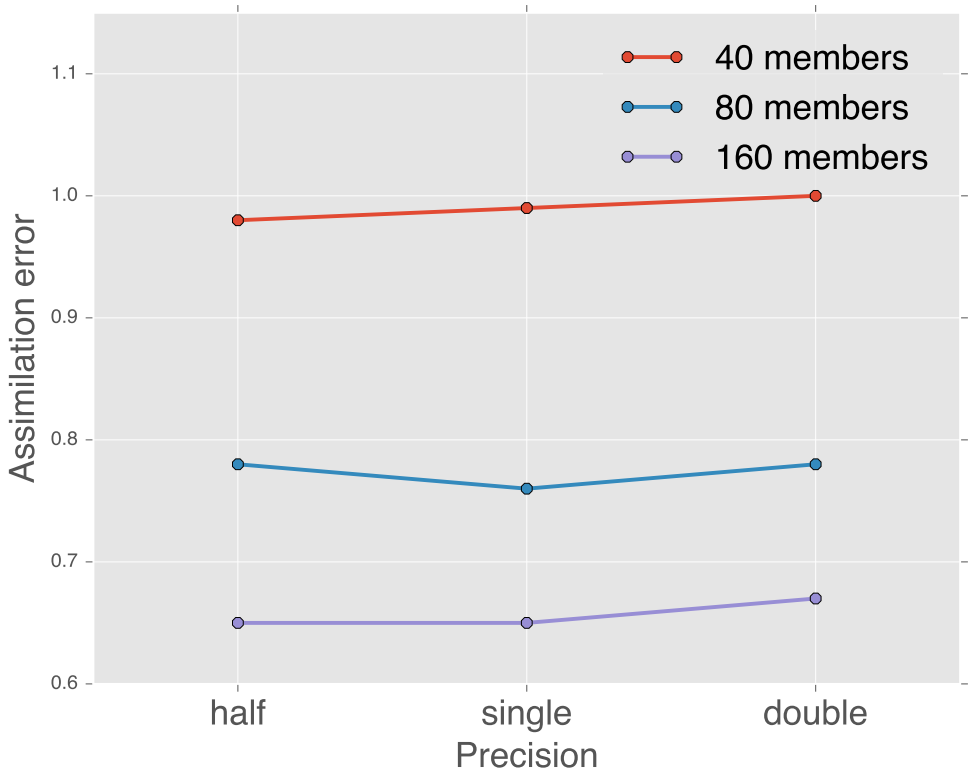
Two research questions



Will our models fail if we reduce precision?

Can we identify the optimal level of precision?

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

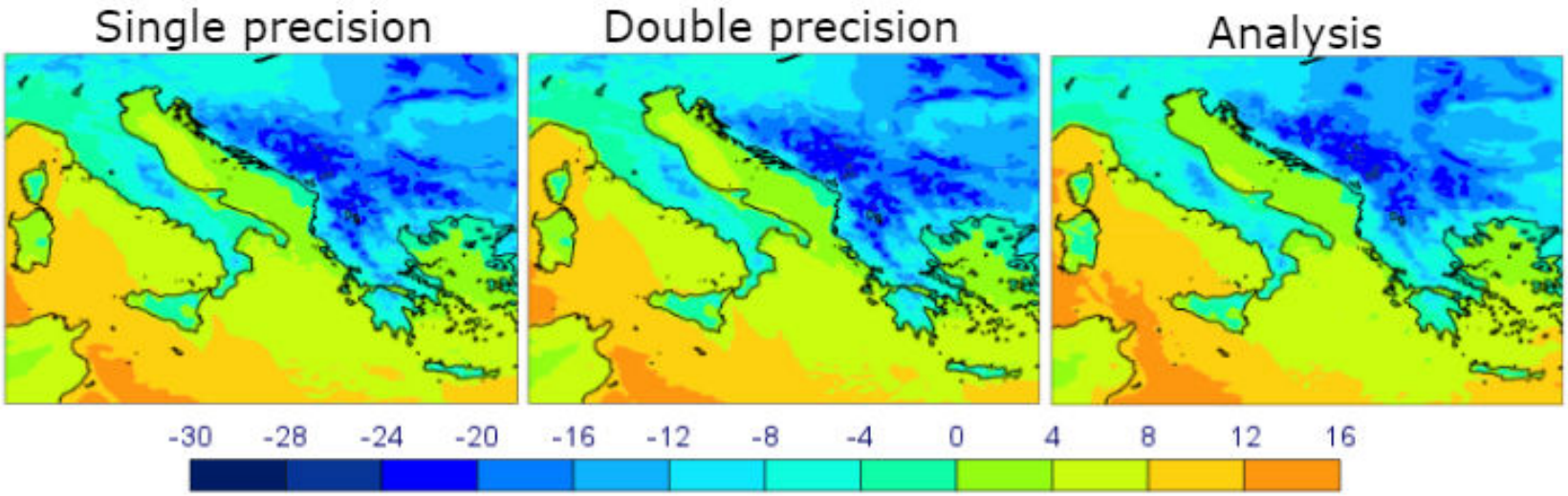
Resolution	Precision in number of bits	Normalised Energy Demand	Mean error Z500 at day 2
235 km	64	1.0	2.3
315 km	64	0.47	4.5
235 km	**20**	0.29	2.5

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.

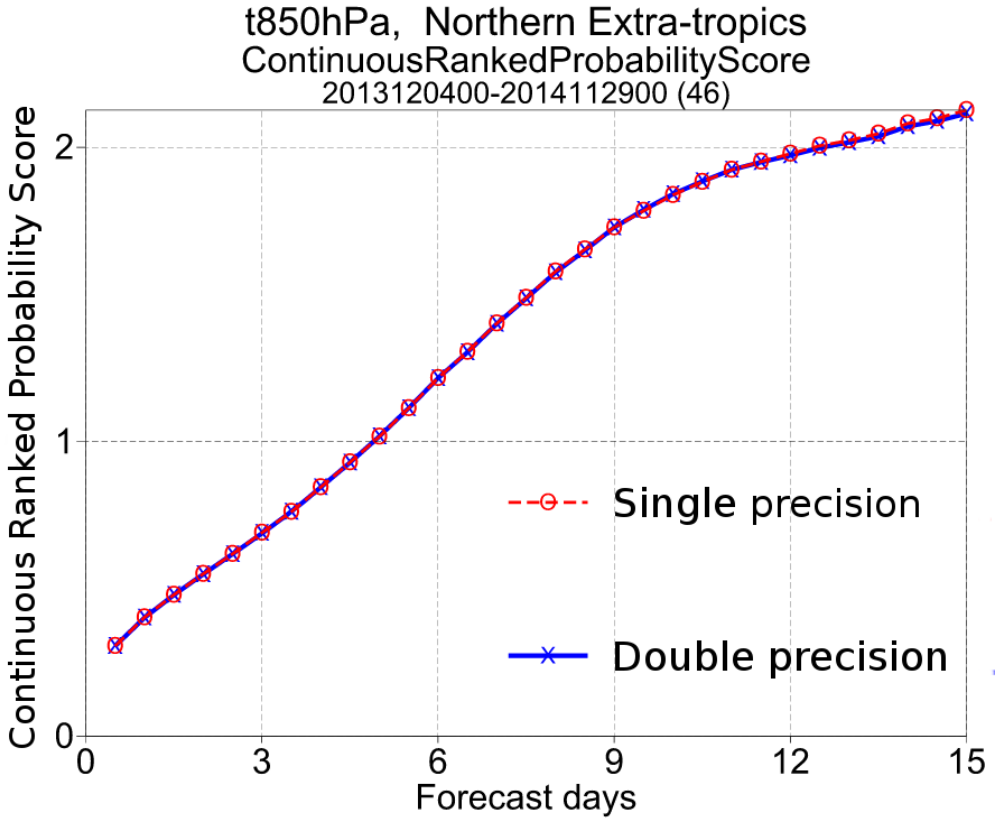
ECMWF's weather forecast model in single precision



- Forecast quality in double and single precision is almost identical.
- 40% reduction of run time.
- Benefit for global simulations at 1.0 km resolution.

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

ECMWF's weather forecast model in single precision

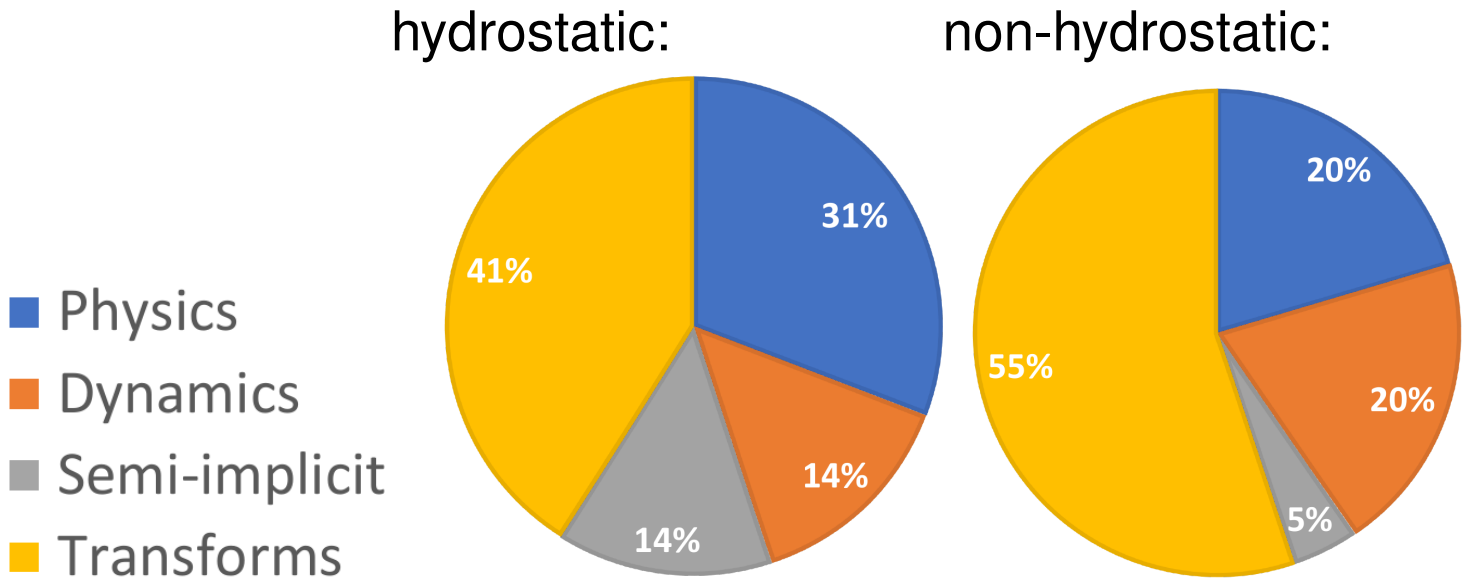


- Forecast quality in double and single precision is almost identical.
- 40% reduction of run time.
- Benefit for global simulations at 1.0 km resolution.

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

Use machine learning hardware

Relative cost for model components at 1.25 km for a spectral model:

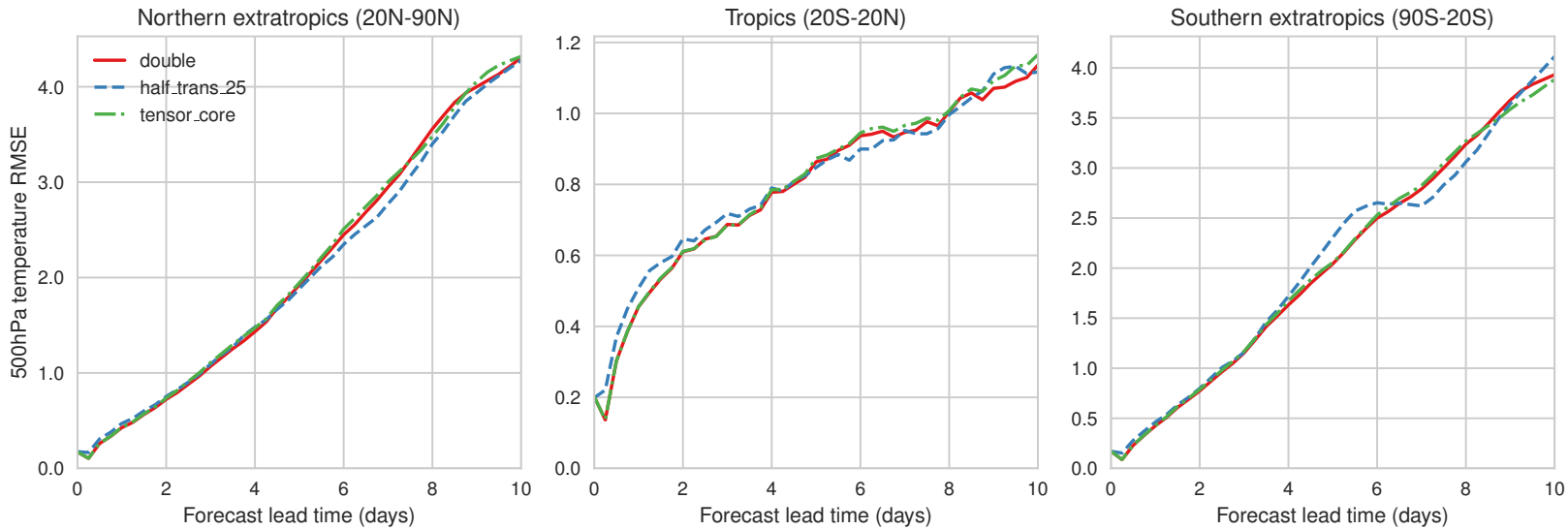


The Legendre transforms are the killer (as expected). They are standard matrix-matrix multiplications.

If we can re-scale the input and output fields, we can use half precision arithmetic (low zonal wave numbers need to be secured).

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



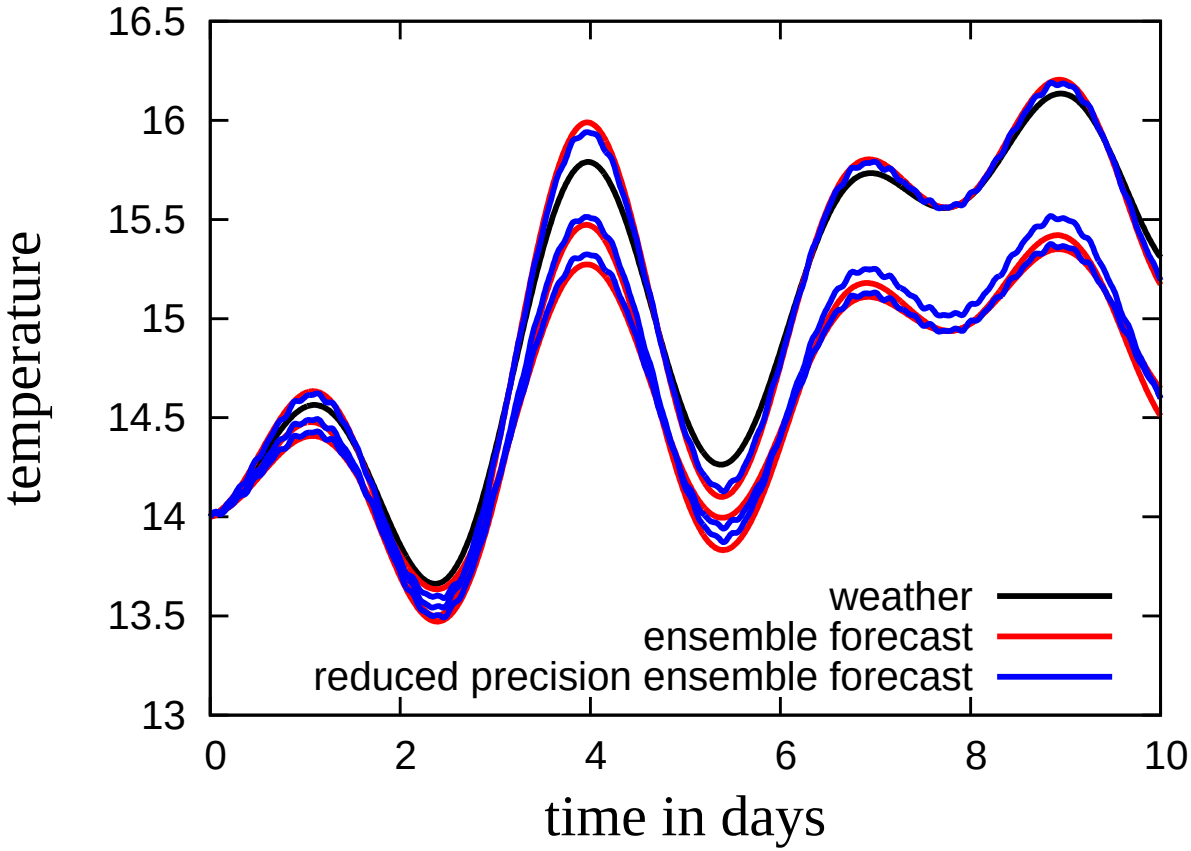
Root-mean-square error for Z500 at TCo1279 resolution averaged over multiple start dates.

Hatfield, Chantry, Dueben, Palmer, submitted to PASC2019.

The simulations are using an emulator to reduce precision.

Dawson and Dueben GMD 2017

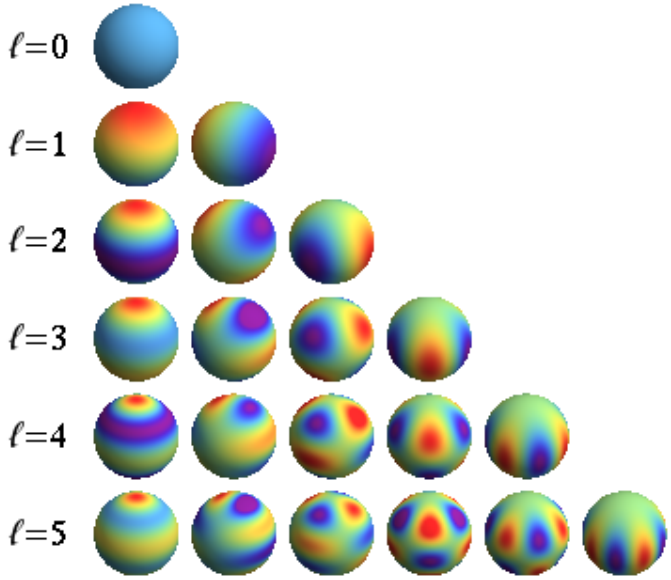
Two research questions



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

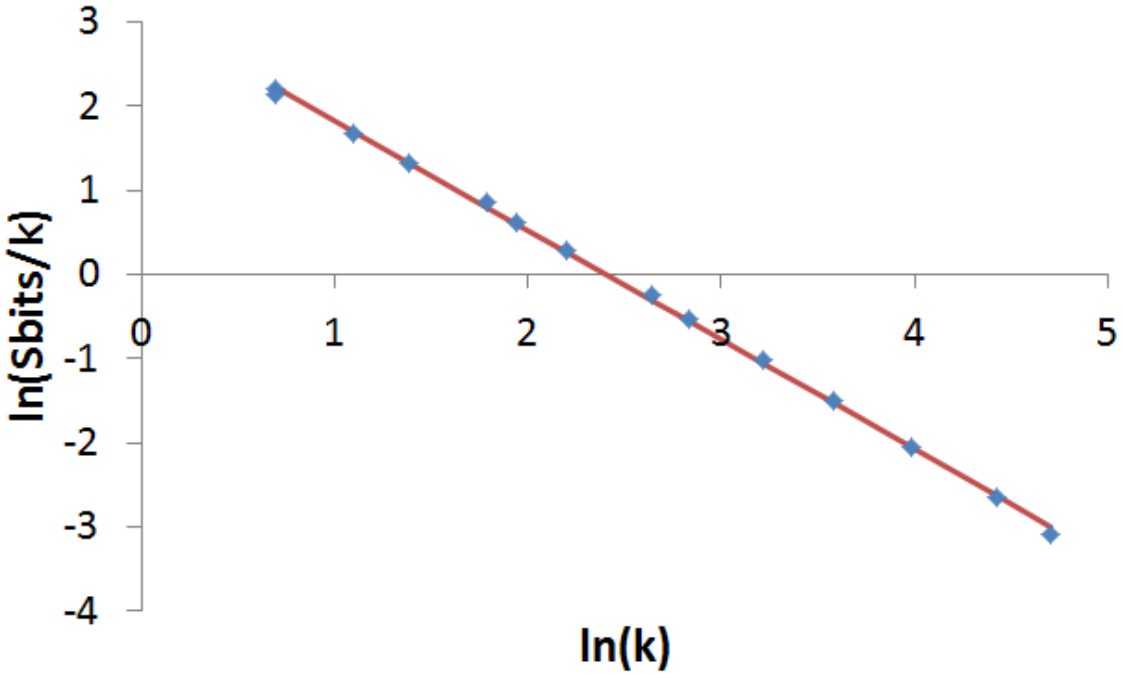
Can we identify the optimal level of precision?

A scale-selective approach



- 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
PhD student Tobias Thornes

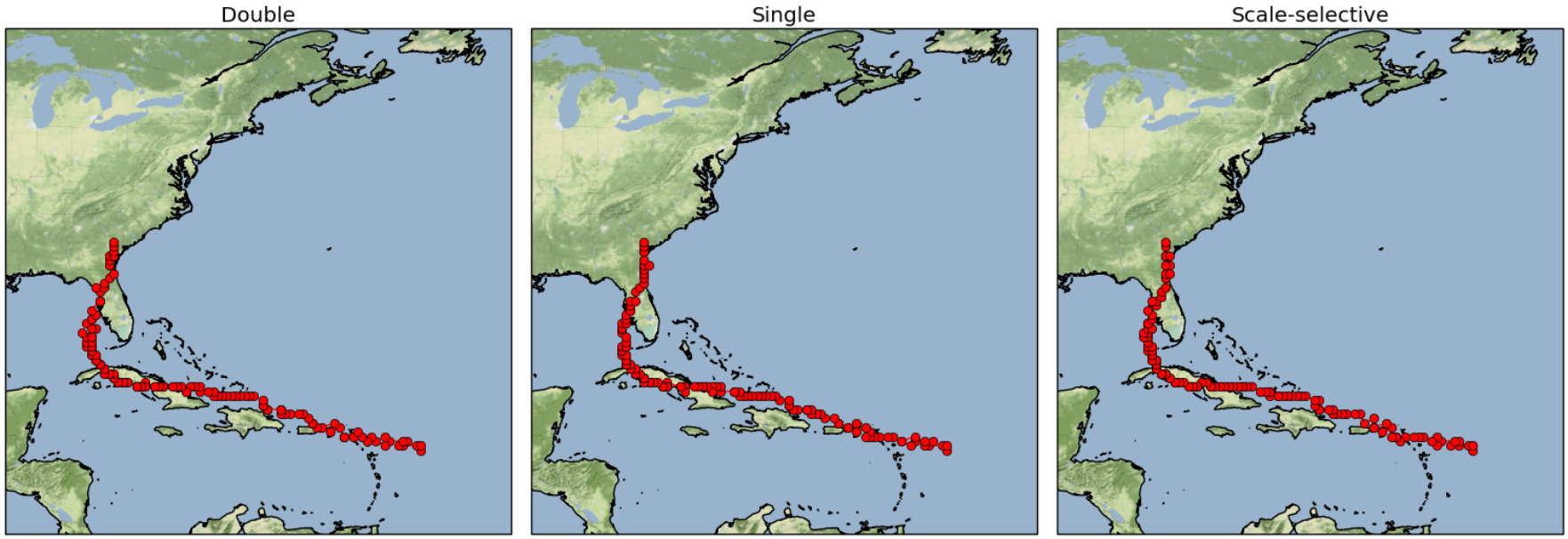


A scale-dependent reduction in precision for the surface quasi-geostropic equations.

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

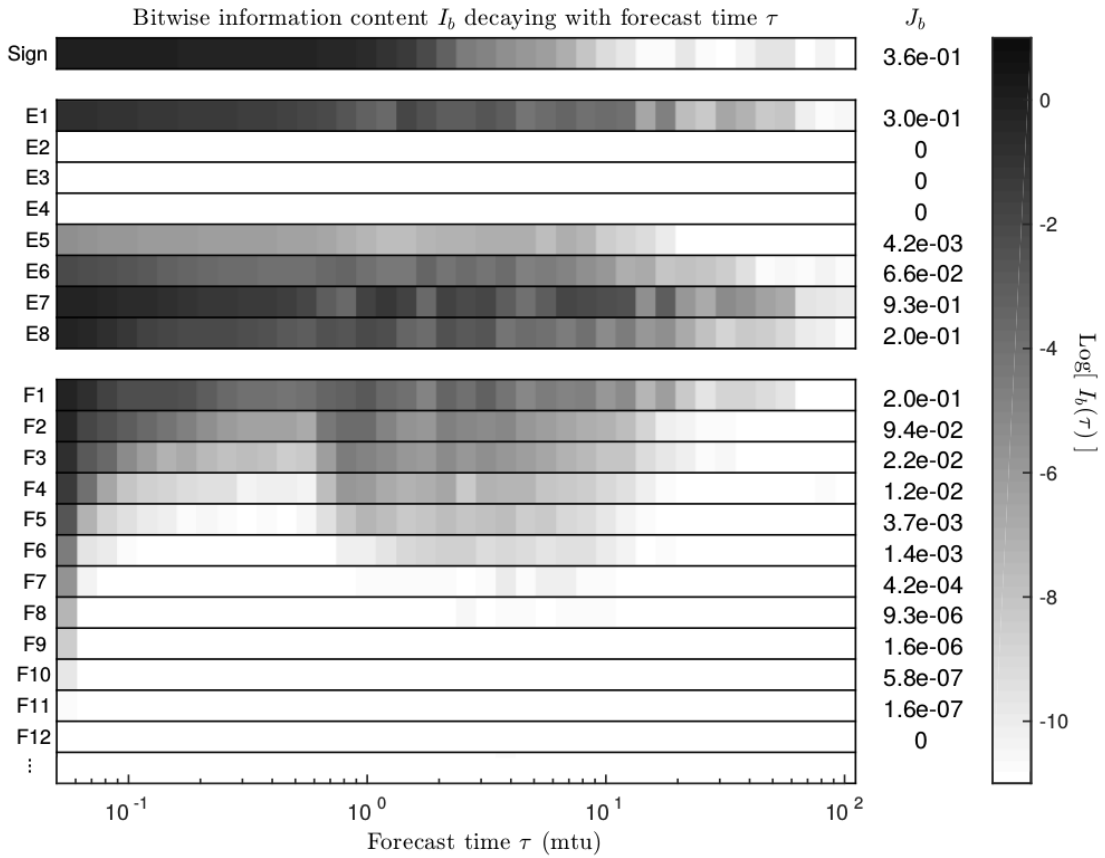
A scale-selective approach: Track of Hurricane Irma PostDoc Matthew Chantry



- 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

Bitwise information content and predictability

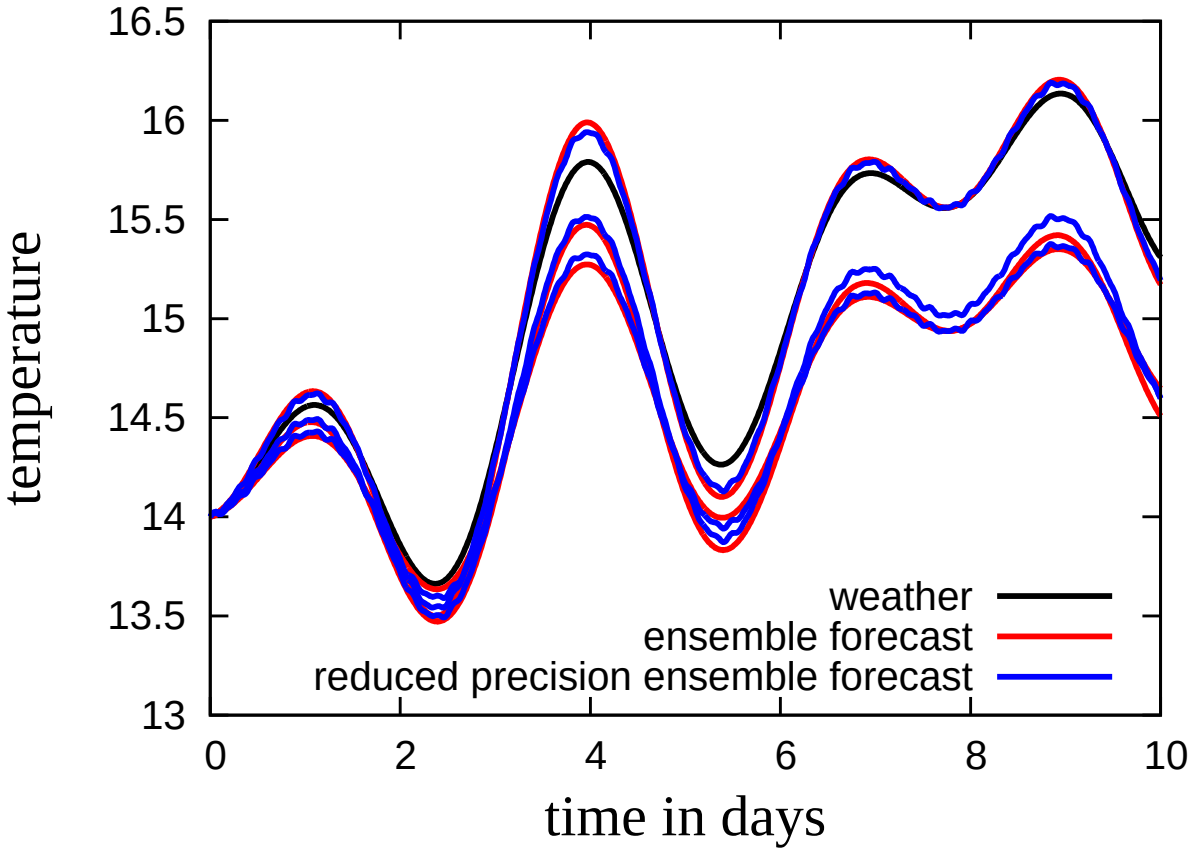


Information content of bits for a Lorenz'63 model using a single long term integration and Shannon information theory.

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

Two research questions



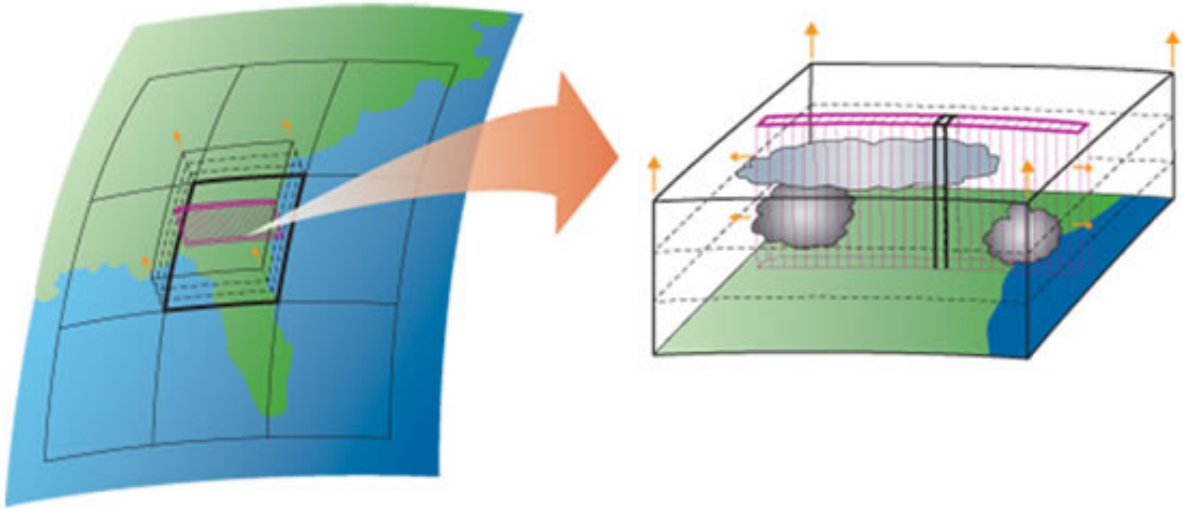
Will our models fail if we reduce precision? - No!

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?

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

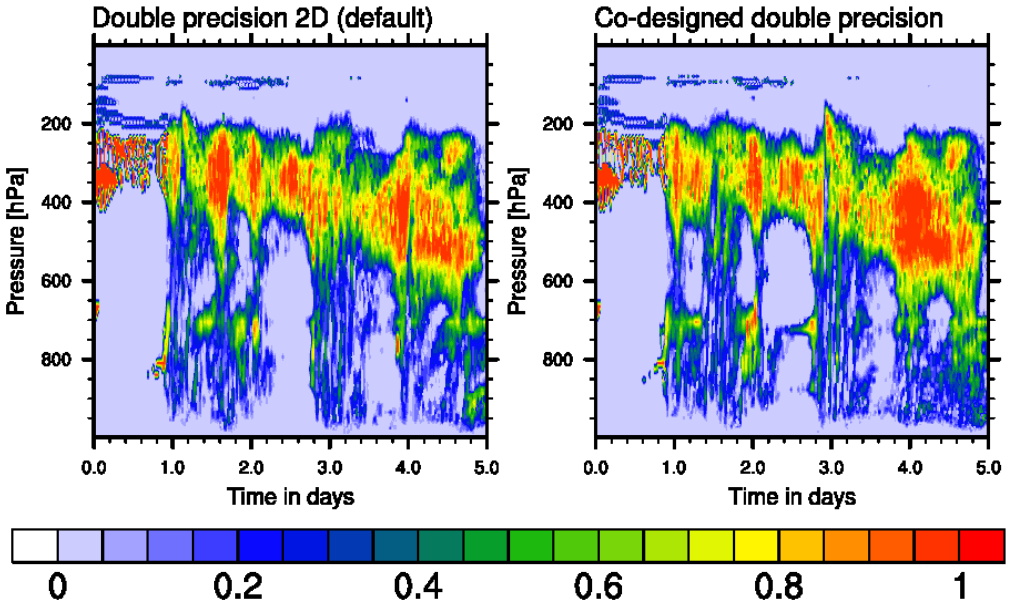
- We automate the search for reduced precision to find the optimal level of precision for individual parameters and model fields.
- We compare model errors due to reduced precision with ensemble spread.

Parameter/Variable	Precision	Relative rounding 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



- Precision can be reduced almost to zero 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.

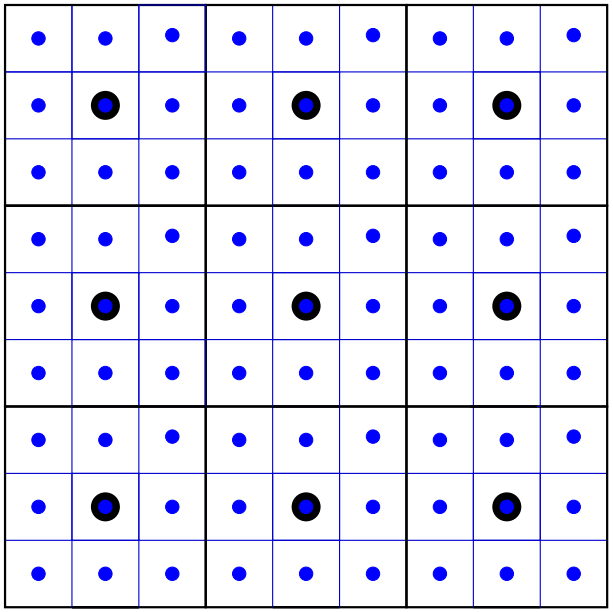
A precision analysis can help to adjust model complexity.

Düben, Subramanian, Dawson and Palmer JAMES 2017

One more research questions

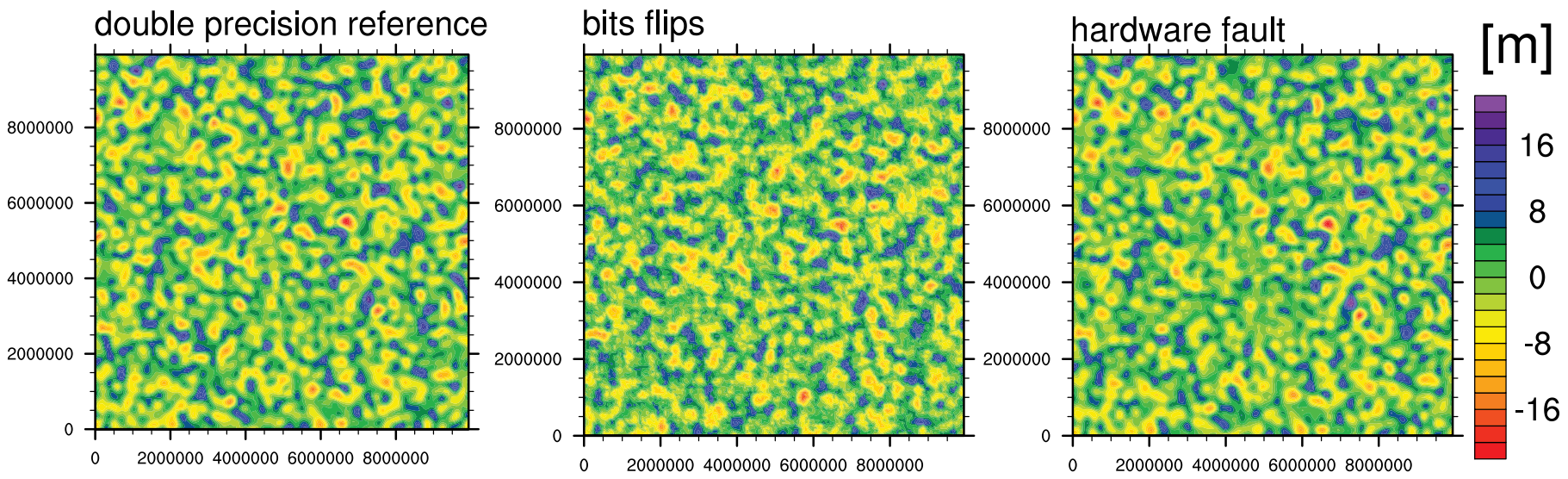
Can a study of numerical precision help to understand model uncertainty and model error?
Yes!

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
```

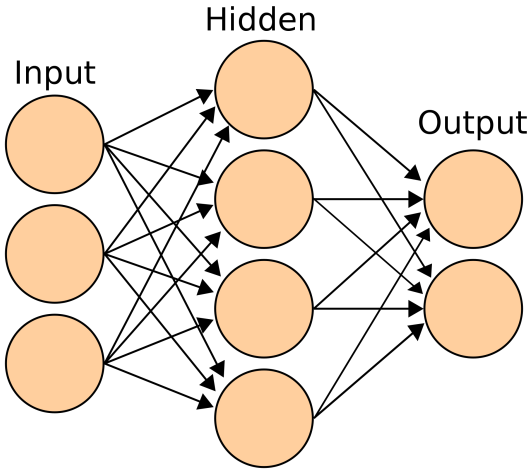
→ *c=3.562500*

Machine learning in weather and climate modelling

- Many techniques can be labelled as “machine learning”.
- We apply machine learning all the time.

- Decision Trees and Random Forests are interesting.
- I will focus on neural networks.

Neural Networks in a nutshell



www.wikipedia.org

- Neural Networks can learn from input/output pairs to emulate a non-linear process.
- Neurons have weighted connections to each other and the weights are trained to produce the optimal results.

In the following, I will show example to (1) emulate existing model components, (2) learn the equations of motion, (3) improve post-processing and (4) use machine learning hardware.

Emulate existing model components

- Store input/output pairs of parametrisation schemes.
- Use this data to train a neural network to do the same job.
- Replace the parametrisation scheme by the neural network.

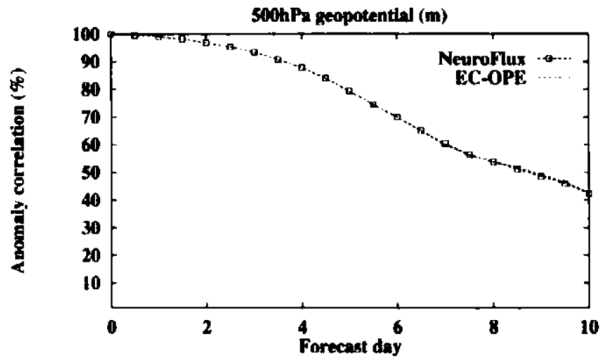
Why would you do this?

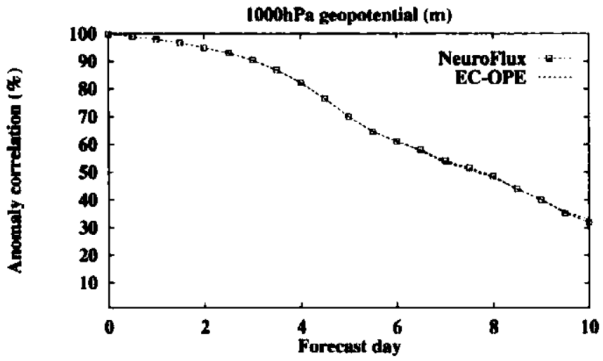
- A large fraction of the computational cost is generated by parametrisation schemes.
- Parametrisation schemes cause $>$ 90% of model code.
- Optimization of this code is very difficult
  ($\rightarrow$ less than 5% peak performance).
- Neural Networks are highly optimized and can even use co-designed hardware.
  $\rightarrow$ Portability comes for free.

We hope that deep Neural Networks will be almost as good as the original parametrisation schemes but much more efficient.

Neural Networks to replace the radiation scheme at ECMWF in the year 2000

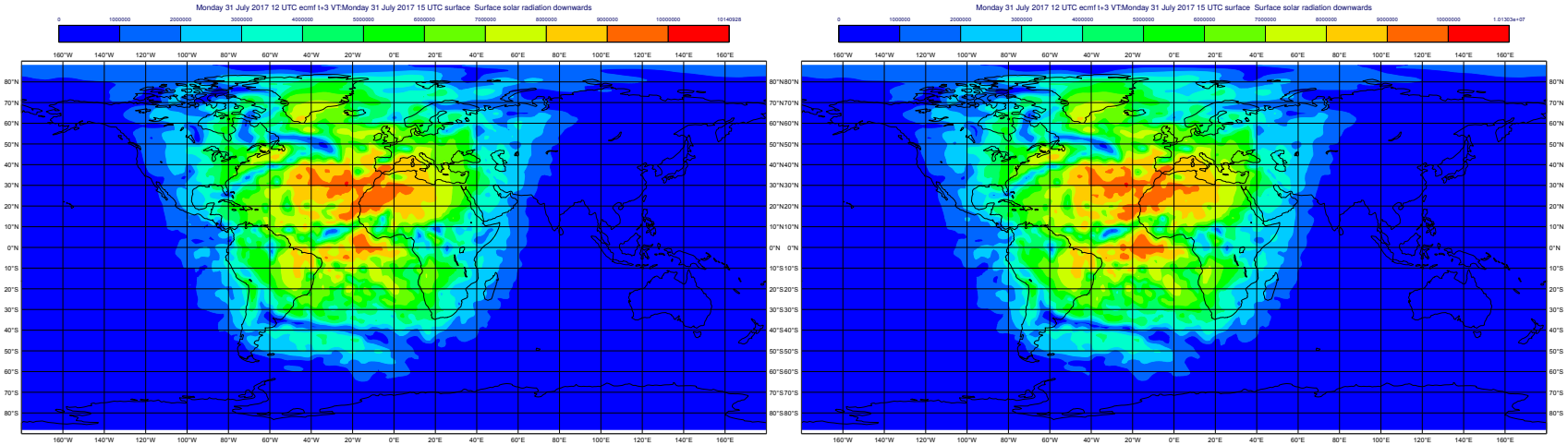
- 20-30 hidden neurons.
- Trained on 80,000 vertical profiles.
- Accuracy of the new scheme was comparable.
- The new scheme was seven times faster.
- The network could be used to generate tangent linear and adjoint code for 4DVar data assimilation.
- However, Neural Networks are currently not used in operational models.





Chevallier et al. QJRMS 2000.

A neural network emulator for the state-of-the-art model configuration with 137 vertical levels

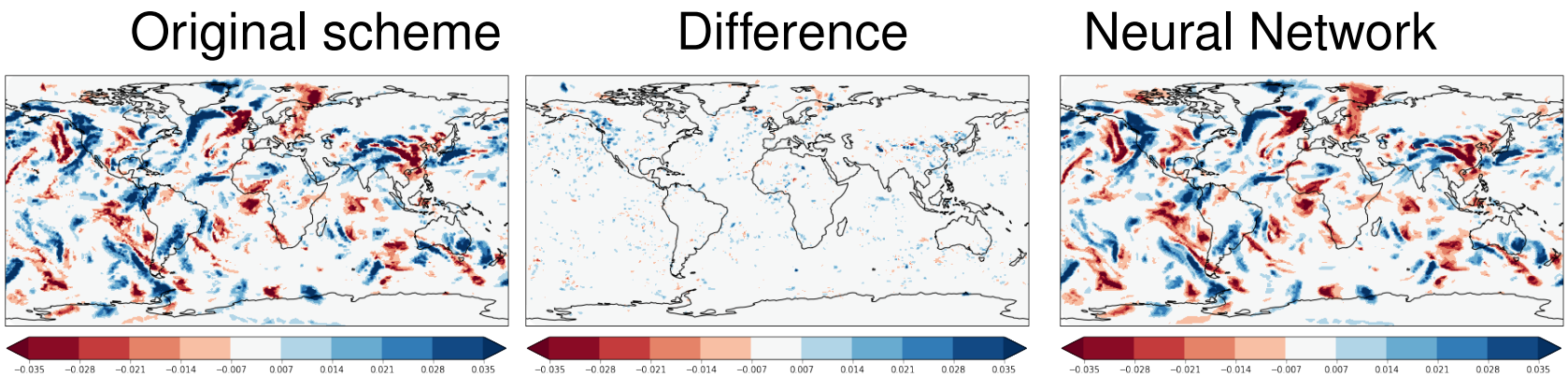


Progsch, Ko, Angerer @NVIDIA and Dueben, Hogan, Bauer @ECMWF

Downward solar radiation at the surface for the original radiation scheme and the Neural Network emulator.

However, we still need to stabilise free-running model simulations with the Neural Network and more work is required.

A neural network emulator for gravity wave drag



Chantry, Abdelrahman, Desai, Dueben, Palem, Palmer.

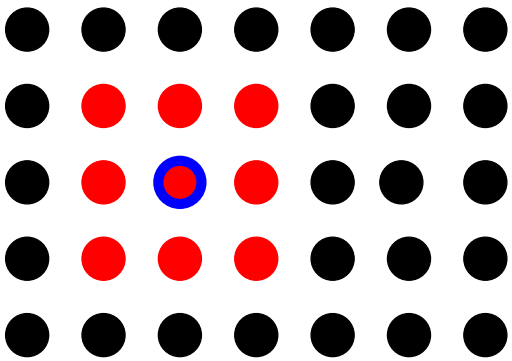
Tendency output for the non-orographic gravity wave drag parametrisation scheme for the standard scheme and a neural network emulator.

Learn the equations of motion

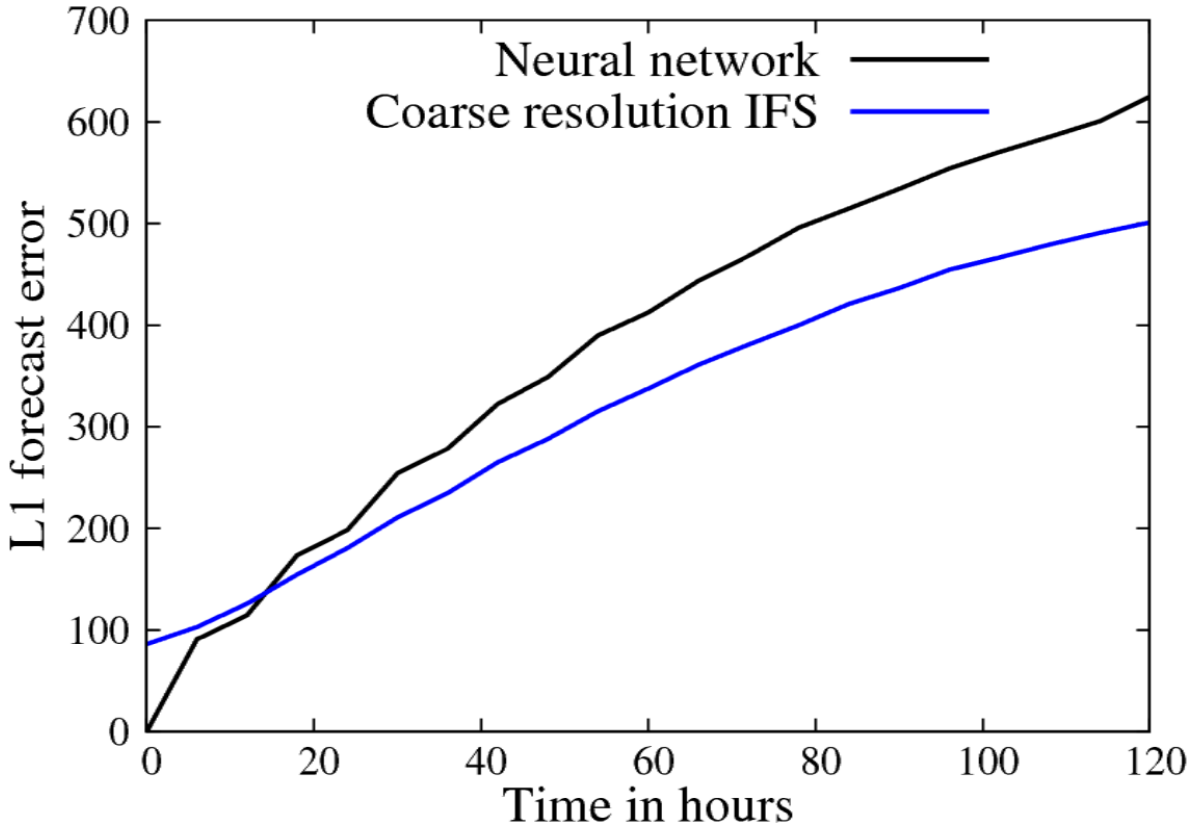
- We know the equations of motion of the atmosphere but we cannot solve them.
- Discretisation and sub-grid-scale variability generates significant errors.
- The data handling system of ECMWF provides access to over 210 petabyte of primary data and the data archive of ECMWF grows by about 233 terabyte per day.

Global weather forecast based on Neural Networks

- Retrieve hourly data of geopotential height at 500 hPa from ERA5 re-analysis for training ($>$ 65000 global data sets).
- Map the data to a coarse lon/lat grid (60x31).
- Use the state of the model at time step i as input and the state of the model at time step $i + 1$ as output.
- Use a 9×9 stencil around the grid point that should be predicted.
- Add time of day and year as well as the coordination of a grid point (lon+lat) as input variables to the network.
- The poles need special treatment.



Global weather forecast based on Neural Networks

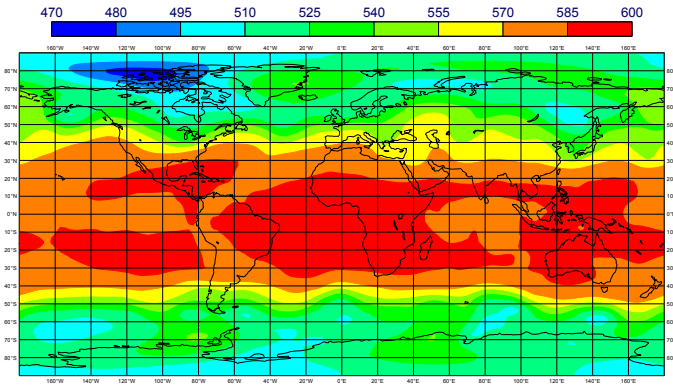


Dueben and Bauer GMD 2018

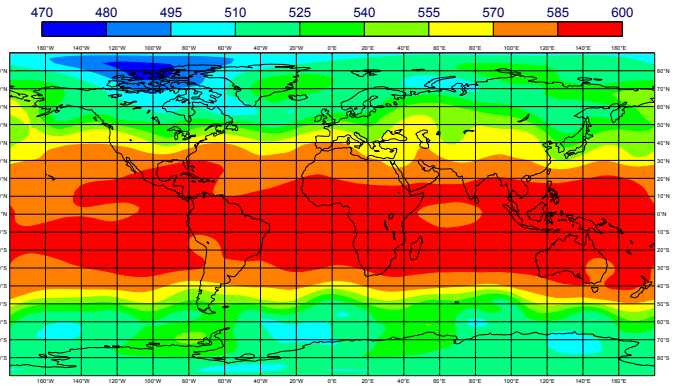
The Neural Network model can compete with a dynamical model of similar complexity.

Global weather forecast based on Neural Networks

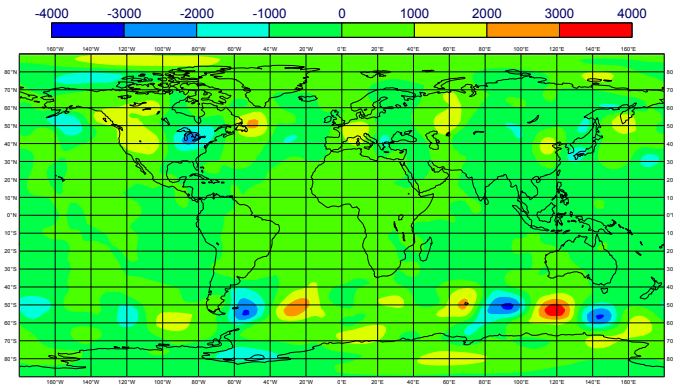
24 hours; Analysis



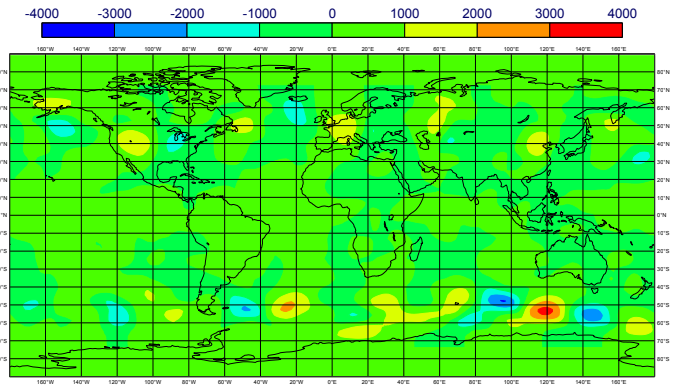
24 hours; Local Neural Network



24 hours; Analysis



24 hours; Local Neural Network



The simulations show reasonable dynamics.

Just adding further inputs does not necessarily help.

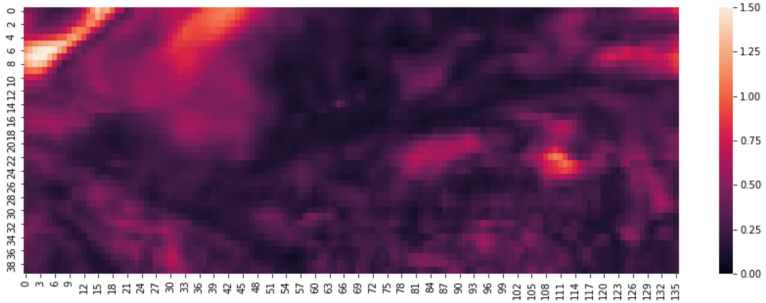
Model runs crash after a couple of weeks.

Improve post-processing

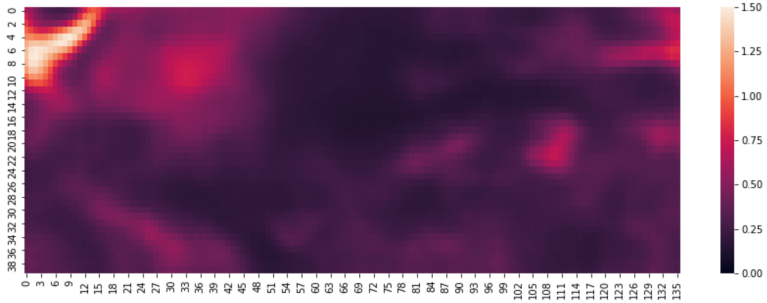
Ensemble simulations are important but expensive.

We use one model trajectory, the ensemble mean and the ensemble spread field at initialisation to predict the ensemble spread of a 10 member ensemble six hours into the forecast for an area over Europe (40W-30E and 40N-60N).

Spread after 6 hours:



Prediction from neural network:



Grönquist, Ben-Nun, Taranov, Höfler @ ETH and Dueben and Bauer @ ECMWF

Challenges 1

Weather and climate models are very complex with non-linear interactions between model components at different time-scales.

There is no fundamental reasons not to use a black box. However,...

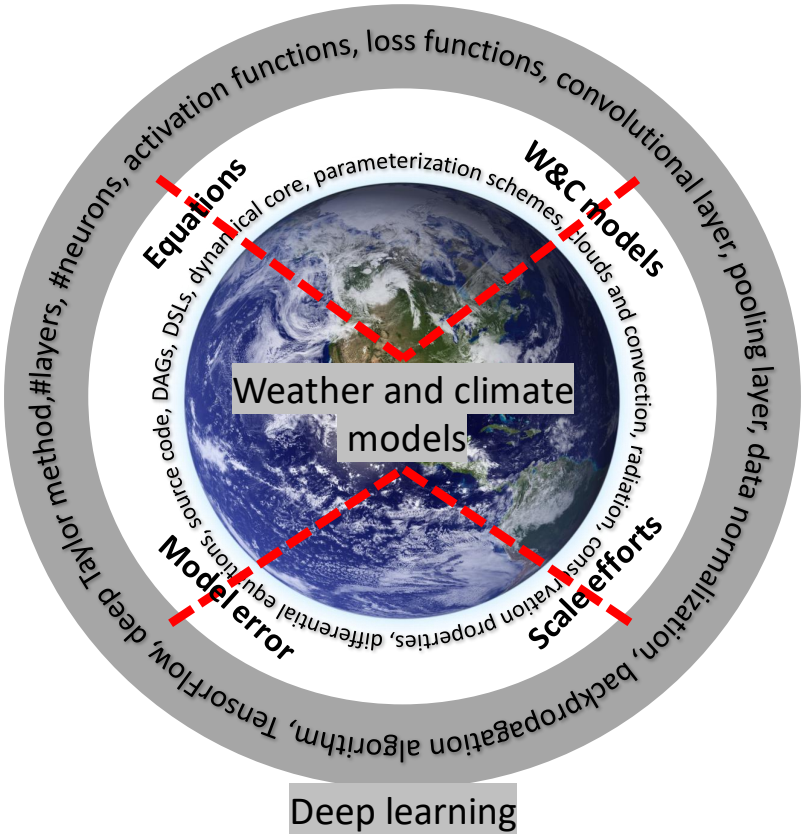
- We have a good knowledge about the Earth System and the leading equations of motion are known for almost all of its components. How can we use this knowledge?
- We do not know how to remove biases via an adjustment of parameters. How shall we deal with this?
- How to adjust fluxes between model components and how to secure conservation laws?
- How to pick hyper-parameters (#neurons, #layers, activation, loss,...) with no use of excessive trial and error testing?
- How can we generate networks that reproduce results if hyper-parameters or training data is changed?

Challenges 2

- How can we get beyond “dense” networks to use scalable methods (convolution, pooling,...)? How to establish the right connectivity between neurons?
- There is no guarantee that the model will interact correctly with the Neural Network parametrisation and the model response may be non-trivial.
- How can we diagnose physical knowledge from the network? How can we “debug” a network?
- How can we stabilize long-term integrations or represent complex interactions between model features.
- Fields are very diverse (specific humidity, precipitation, geopotential height, surface pressure,...).
- What can be used as a “better” truth? Superparametrisation, Large Eddy Simulations, high-resolution simulations.
- Can a Neural Network parametrisation scheme explore the full phase space (all weather regimes) during training?

The way forward

- To study known differential equations to learn how to derive blueprints for neural network architectures.
- To study model source code to learn how to derive blueprints for the design of network architectures.
- To study the representation of sub-grid-scale processes and systematic errors when using neural networks.
- To scale the application of neural networks in W&C models beyond today's limits.



An example: The Burgers equation

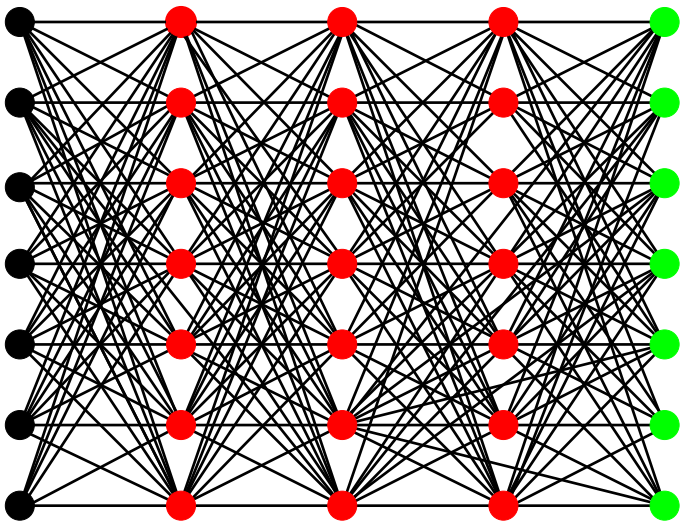
Let's represent a non-linear system that is approximated by the Burgers' equation:

$$\frac{\partial u}{\partial t} = \nu \frac{\partial^2 u}{\partial x^2} - u \frac{\partial u}{\partial x} + p.$$

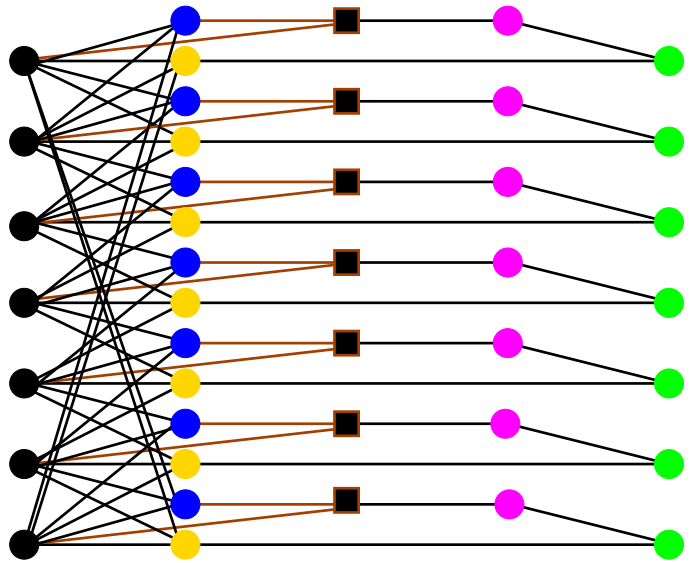
The conventional approach:

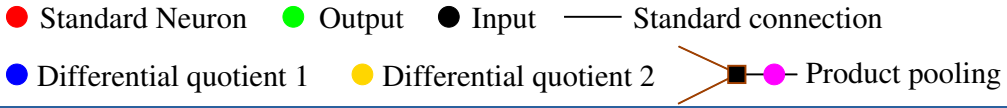
$$\frac{\partial u_i}{\partial t} = \nu \frac{u_{i+1} - 2u_i + u_{i-1}}{\Delta x^2} - u_i \frac{u_{i+1} - u_{i-1}}{2\Delta x} + c_0 + c_1 \cdot u_i + c_2 \cdot u_i^2 + c_3 \cdot u_i \cdot \zeta.$$

The data-science approach:



The way forward:





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 will allow significant savings.
- savings can be reinvested to achieve higher resolution/complexity or more ensemble members to improve predictions.
- our understanding of model error and model uncertainty helps to adjust precision.
- machine learning may open up opportunities to increase efficiency and improve models.