# More Accuracy with Less Precision:
## An Information Theoretic Paradigm for Weather and Climate Simulation

Tim Palmer
Department of Physics, University of Oxford

With thanks to: Andrew Dawson, Peter Düben, Stephen Jeffress, Dave Macleod

# Strategy

**Update July 2016: The Council has approved the new strategy for the period 2016–2025.**

## Goals by 2025

To provide forecast information needed to help save lives, protect infrastructure and promote economic development in Member and Co-operating States through:

**Research** at the frontiers of knowledge to develop an integrated global model of the Earth system to produce forecasts with increasing fidelity on time ranges up to one year ahead. This will tackle the most difficult problems in numerical weather prediction such as the currently low level of predictive skill of European weather for a month ahead.

**Operational ensemble-based analyses and predictions** that describe the range of possible scenarios and their likelihood of occurrence and that raise the international bar for quality and operational reliability. Skill in medium-range weather predictions in 2016, on average, extends to about one week ahead. By 2025 the goal is to make skilful ensemble predictions of high-impact weather up to two weeks ahead. By developing a seamless approach, we also aim to predict large-scale patterns and regime transitions up to four weeks ahead, and global-scale anomalies up to a year ahead.

TECHNICAL MEMORANDUM

598

Stochastic Parametrization and Model Uncertainty

Palmer, T.N., R. Buizza, F. Doblas-Reyes, T. Jung, M. Leutbecher, G.J. Shutts, M. Steinheimer, A. Weisheimer

Research Department

October 8, 2009



European Centre for Medium-Range Weather Forecasts
Europäisches Zentrum für mittelfristige Wettervorhersage
Centre européen pour les prévisions météorologiques à moyen terme

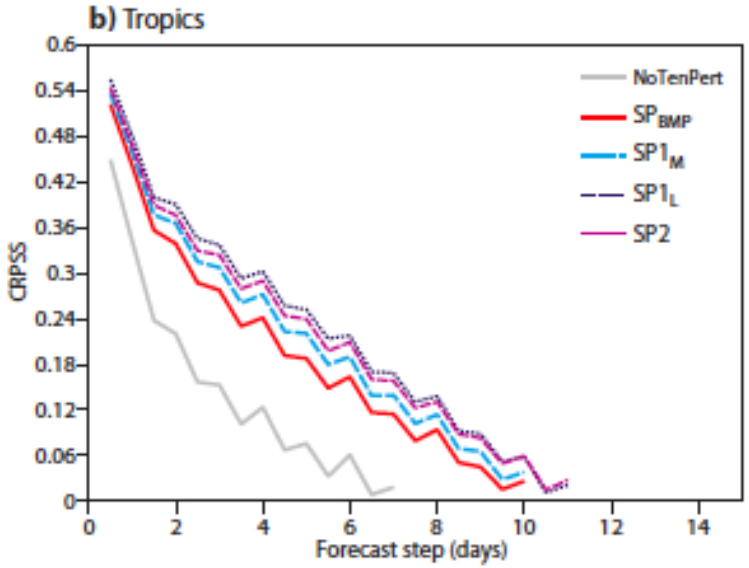
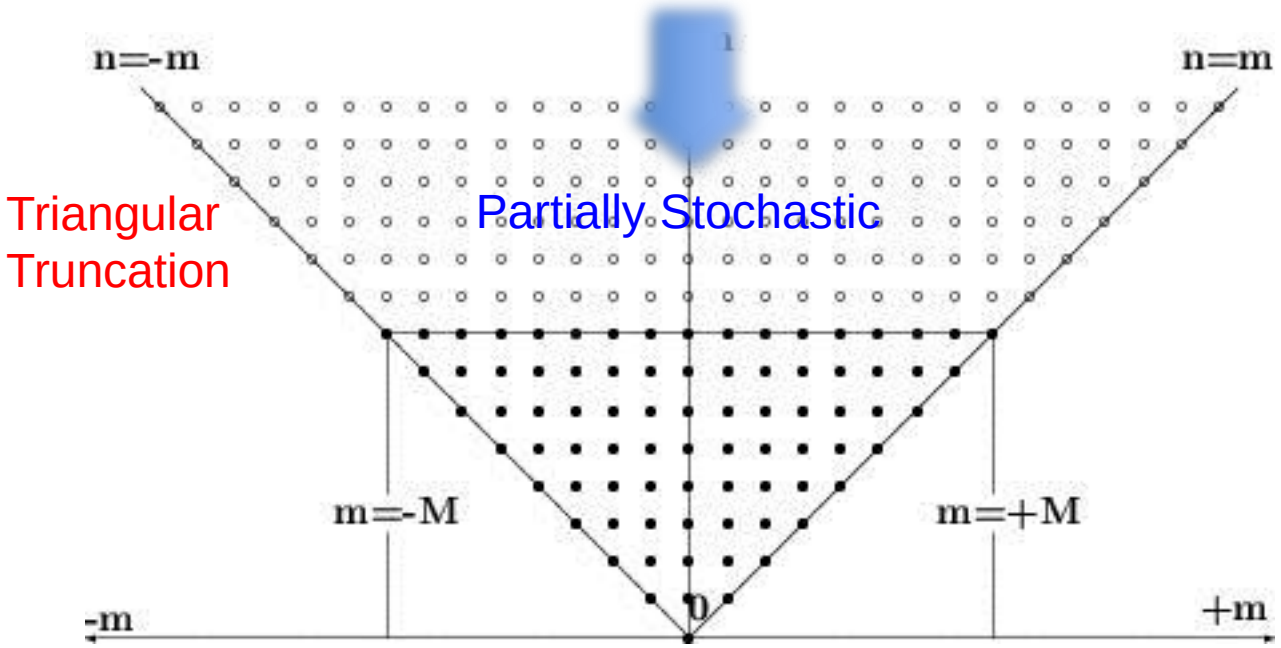



*Figure 3: Continuous Ranked Probability Skill Score for 850 hPa temperature.*



Irreducible Uncertainty in sub-grid representations

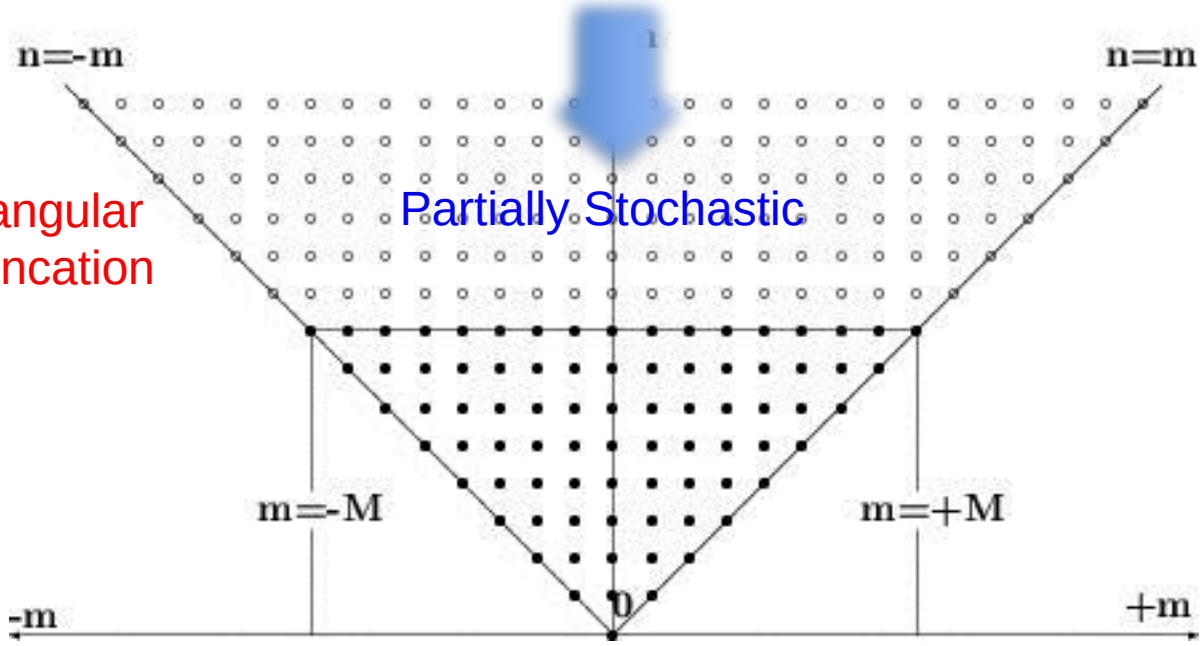
# Stochastic Parametrisation




If parametrisation is partially stochastic, are we “over-engineering” our dynamical cores by using 64-bit precision for all variables?

Are we making inefficient use of computing resources (i.e. energy) that could otherwise be used to increase resolution towards convective scales?

Stochastic Parametrisation/
Earth System Complexity




Triangular Truncation

Partially Stochastic

Quarter precision?

Half precision?

Single Precision?

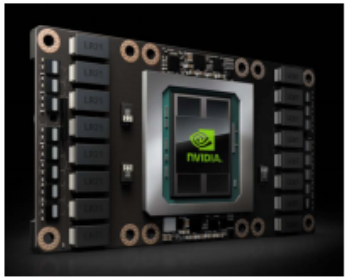
Double precision

What is the real information content in each of the billions of bits that represent variables in a weather/climate model?

April 5, 2016

## NVIDIA Unleashes Monster Pascal GPU Card at GTC16

Tiffany Trader




Earlier today (Tuesday) at the seventh-annual GPU Technology Conference (GTC) in San Jose, Calif., NVIDIA revealed its first Pascal-architecture based GPU card, the P100, calling it "the most advanced accelerator ever built." The P100 is based on the NVIDIA Pascal GP100 GPU — a successor to the Kepler GK110/210 — and is aimed squarely at HPC, technical computing and deep learning workloads.

Packing a whopping 5.3 teraflops of double-precision floating point performance, the P100 is NVIDIA's most performant chip to date. And with 15.3 billion transistors, it's also the largest GPU that NVIDIA has ever made in spite of it being built on TSMC's 16nm FinFET manufacturing process.

April 6, 2016

## Europe's Fastest Supercomputer to Get Pascal GPU Upgrade

Tiffany Trader and John Russell

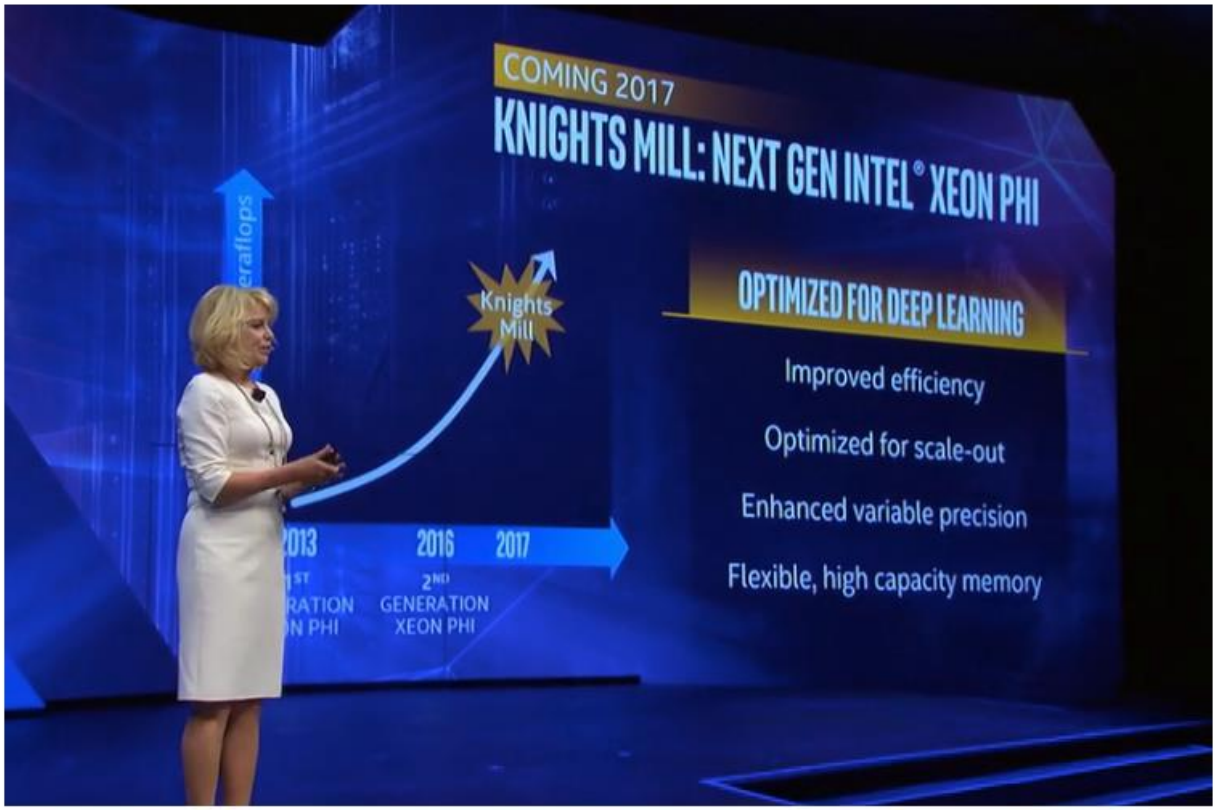

Already Europe's fastest supercomputer at 7.8 petaflops, the Piz Daint (hybrid CPU/GPU Cray XC30) at the Swiss National Computing Center (CSCS) will double its performance with a massive upgrade that involves switching to NVIDIA's newest Pascal GPU architecture and merging with Piz Dora (Cray XC40), a smaller CPU-based machine. The announcement was made at GTC16 yesterday. Last November Piz Daint placed seventh on the TOP500 list.

Plans call for 5,200 NVIDIA K20xs to be replaced by 4,500 Pascal GPUs – which version hasn't been decided. Also, the Intel processors will be upgraded from Sandy Bridge to Haswell architecture. When completed, the new combined system, all on a single fabric, will keep the Piz Daint name and provide

# Intel Unveils Plans for Knights Mill, a Xeon Phi for Deep Learning

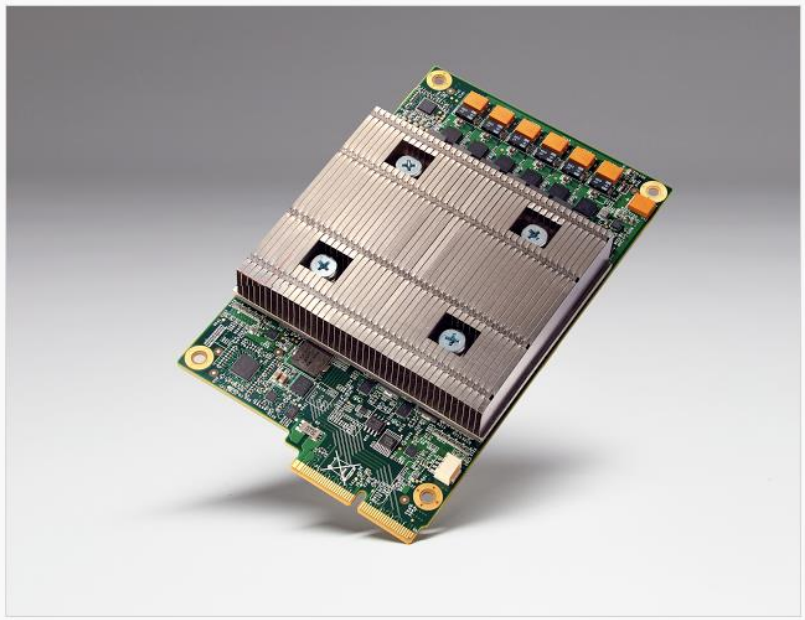
**Michael Feldman, Aug. 18, 2016, 1:33 a.m.**

At the Intel Developer Forum (IDF) this week in San Francisco, Intel revealed it is working on a new Xeon Phi processor aimed at deep learning applications. Diane Bryant, executive VP and GM of Intel's Data Center Group, unveiled the new chip, known as Knights Mill, during her IDF keynote address on Wednesday.



# Google built its own chips to expedite its machine learning algorithms

*Posted May 18, 2016 by* ***Frederic Lardinois*** *(@fredericl)*



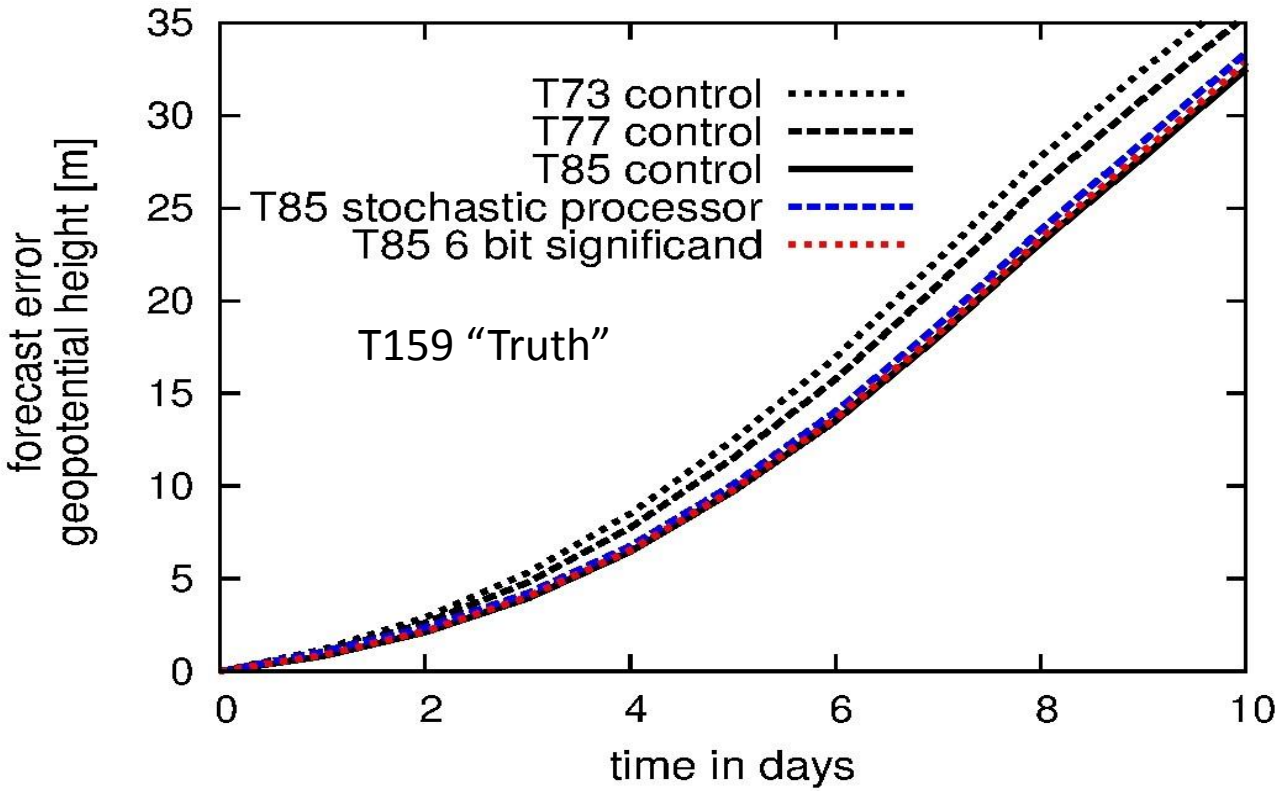
These so-called Tensor Processing Units (TPU) are custom-built chips that Google has now been using in its own data centers for almost a year, as Google's senior VP for its technical infrastructure Urs Holzle noted in a press conference at I/O. Google says it's getting "an order of magnitude better-optimized performance per watt for machine learning" and argues that this is "roughly equivalent to fast-forwarding technology about seven years into the future."

Google also manages to speed up the machine learning algorithms with the TPUs because it doesn't need the high-precision of standard CPUs and GPUs. Instead of 32-bit precision, the algorithms happily run with a reduced precision of 8 bits, so every transaction needs fewer transistors.

If you are using Google's voice

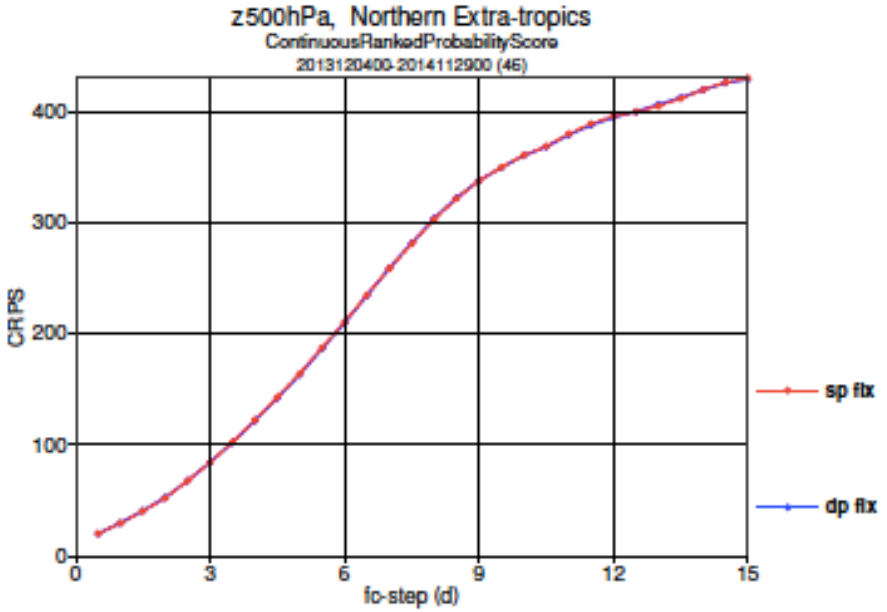
# More accurate “weather forecasts“ with less precision
# Reading Spectral Model

Düben and Palmer, 2014. Monthly Weather Review




The stochastic chip / reduced precision emulator is used on 50% of numerical workload:
All floating point operations in grid point space
All floating point operations in the Legendre transforms between wavenumbers 31 and 85.
T85 cost approx that of T73

Vana et al MWR 2016



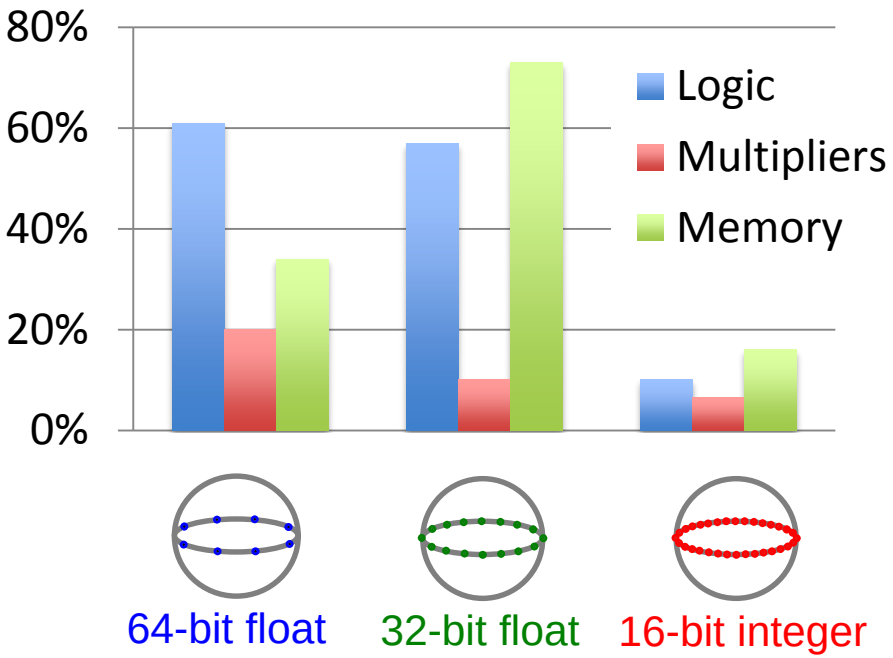

Single vs Double Precision in IFS

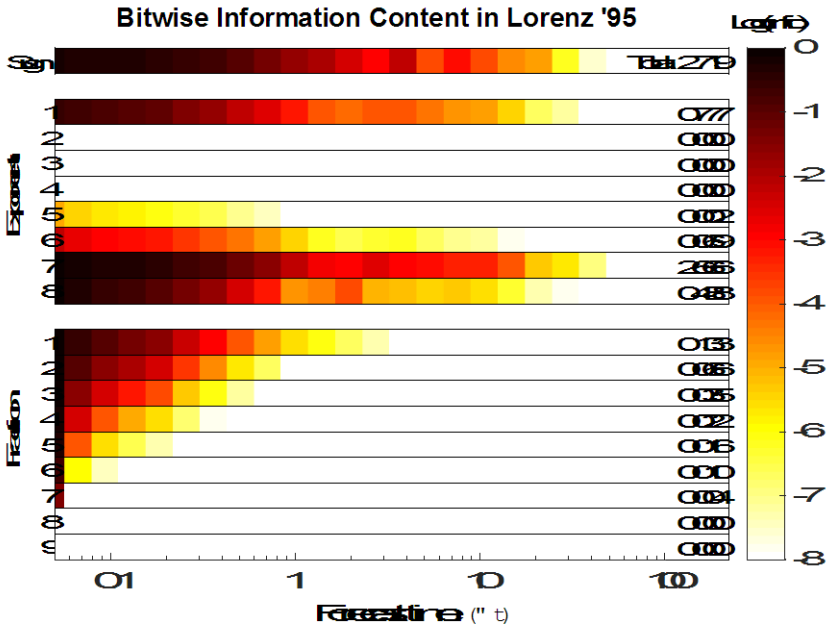
# Maxeler FPGA

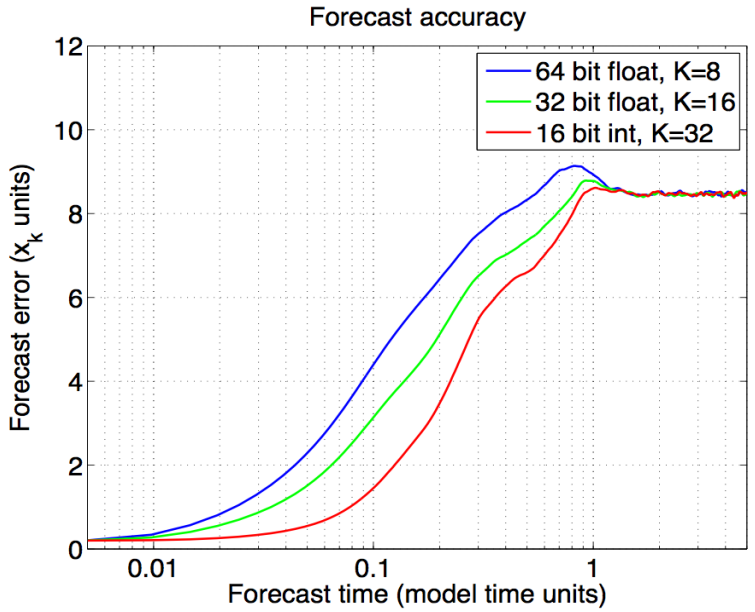
$$\dot{X}_k = X_{k-1}(X_{k+1} - X_{k-2}) - X_k + F$$

Bitwise information content:

$$J_b = \int_0^{\infty} I_b(\Delta t)\, d\Delta t$$



# Tests with a C-grid shallow-water model

| Parameter | number of bits in the significand |
|---|---|
| $\eta$ | 12 |
| u | 12 |
| v | 12 |
| $\tau_x$ | 2 |
| $\tau_y$ | 2 |
| dh | 2 |
| du | 2 |
| dv | 2 |
| ab | 2 |
| h0 | 2 |
| fu | 4 |
| fv | 4 |
| b | 7 |
| zeta | 2 |
| vec | 2 |
| g | 5 |
| pi | 2 |
| f0 | 2 |
| beta | 2 |
| $\nu$ | 2 |
| ah | 2 |
| x0 | 2 |
| y0 | 2 |
| dx | 8 |
| dy | 8 |
| dt | 2 |
| slip | 2 |
| sigmax | 2 |
| sigmay | 2 |

We end up with precision levels that should be used for the significand of floating point numbers.

Precision can be reduced significantly!

We obtain information on the information content.

Expert knowledge is needed to obtain stable model simulations (increase precision for $\Delta t$ and ab).

# 8-bit chips for weather/climate. Completely Mad?

Suppose the dynamic range of a prognostic variable X is $10^{-10}$ to $10^{10}$ , but the dynamics are particularly uncertain (many Earth-System components).

Why not transform the prognostic variable to log X ?

Clim Dyn (2016) 46:3865–3882
DOI 10.1007/s00382-015-2809-5

# The reliability of single precision computations in the simulation of deep soil heat diffusion in a land surface model

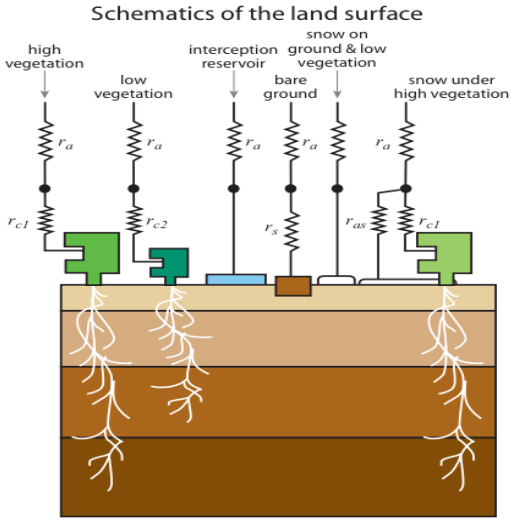
**Richard Harvey**[1,2] · **Diana L. Verseghy**[1]

**Abstract** Climate models need discretized numerical algorithms and finite precision arithmetic to solve their differential equations. Most efforts to date have focused on reducing truncation errors due to discretization effects, whereas rounding errors due to the use of floating-point arithmetic have received little attention. However, there are increasing concerns about more frequent occurrences of rounding errors in larger parallel computing platforms (due to the conflicting needs of stability and accuracy vs. performance), and while this has not been the norm in climate and forecast models using double precision, this could change with some models that are now compiled with single precision, which raises questions about the validity of using such low precision in climate applications. For example, processes occurring over large time scales such as permafrost thawing are potentially more vulnerable to this issue. In this study we analyze the theoretical and experimental effects of using single and double precision on simulated deep soil temperature from the Canadian LAnd Surface Scheme (CLASS), a state-of-the-art land surface model. We found that reliable single precision temperatures are limited to depths of less than about 20–25 m while double precision shows no loss of accuracy to depths of at least several hundred meters. We also found that, for a given precision level, model accuracy *deteriorates* when using *smaller* time steps, further reducing the usefulness of single precision. There is thus a clear danger of using single precision in some climate model applications, in particular any scientifically meaningful study of deep soil permafrost must at least use double precision. In addition, climate modelling teams might well benefit from paying more attention to numerical precision and roundoff issues to offset the potentially more frequent numerical anomalies in future large-scale parallel climate applications.

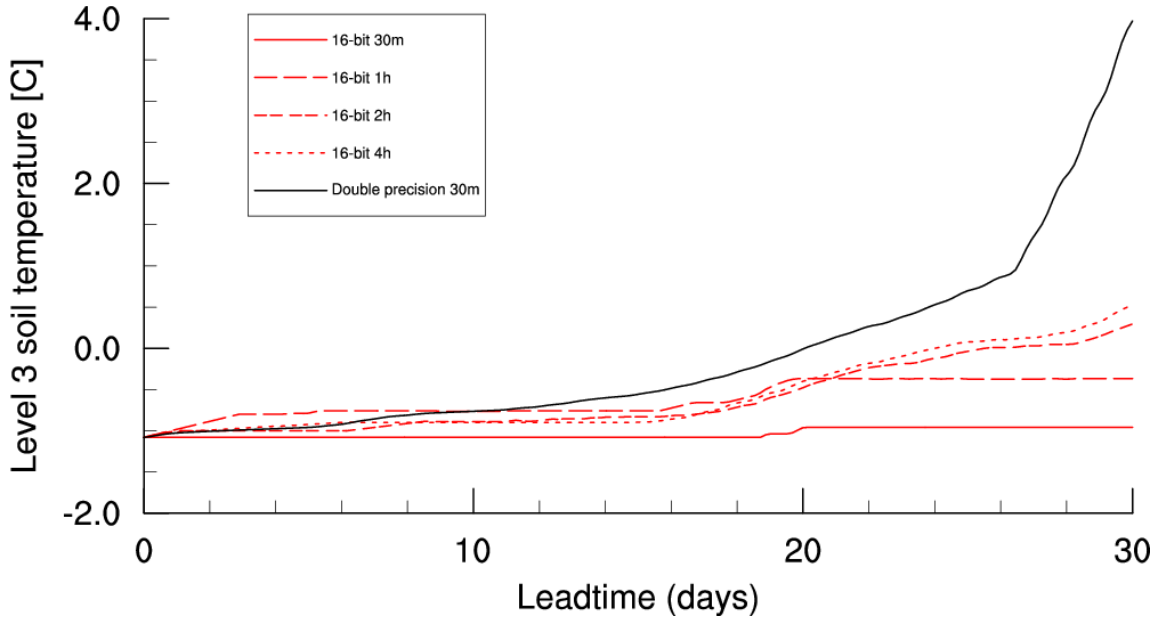
**Keywords** Floating-point arithmetic · Numerical precision · Single precision arithmetic · Double precision arithmetic · Climate models · Permafrost · Land surface models · Deep soil processes

## 1 Introduction

Climate models use sophisticated numerical algorithms to solve the complex primitive equations of atmospheric and oceanic motions. These algorithms contain two well-known and unavoidable sources of errors: *truncation errors* (because computations must be completed in a finite time), which are caused by replacing the continuous time and space differentials of the original field equations with finite increments, and *rounding errors* (because computer memory is not infinite), which are caused by replacing real numbers of infinite precision with finite-sized com-




Dave Macleod, Peter Düben, Andrew Dawson



Highly uncertain

$$\frac{\P y}{\P t} = S(y, f, t)$$

$$y(t^{n+1}) = y(t^n) + \Delta t \cdot S(y(t^n), f(t^n), t^n)$$

Represent at high precision

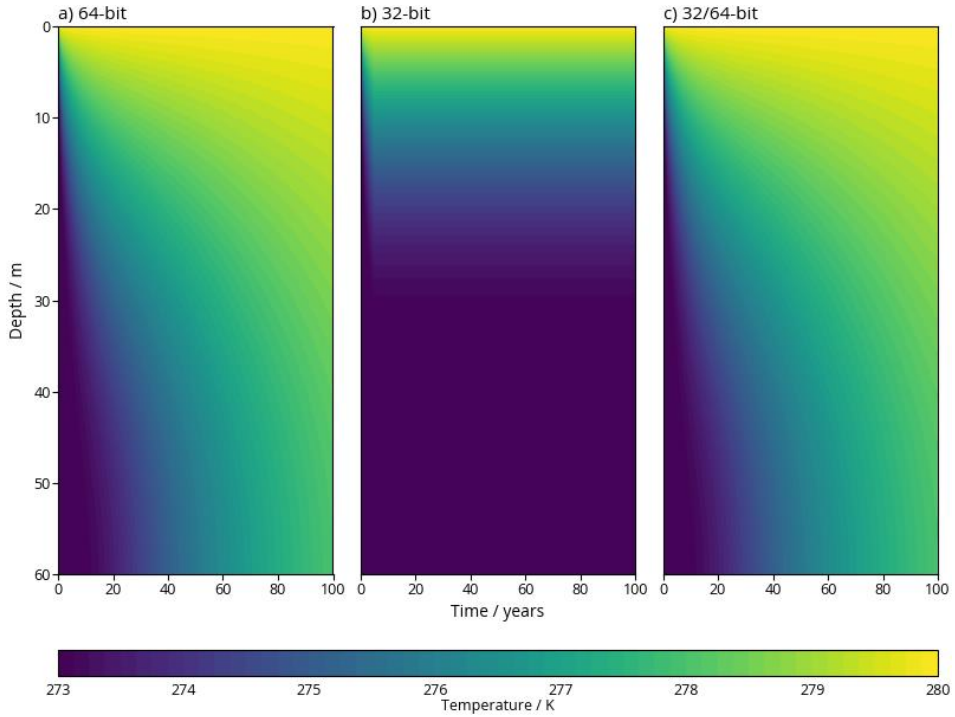
Compute (and retrieve fields from memory) at low precision

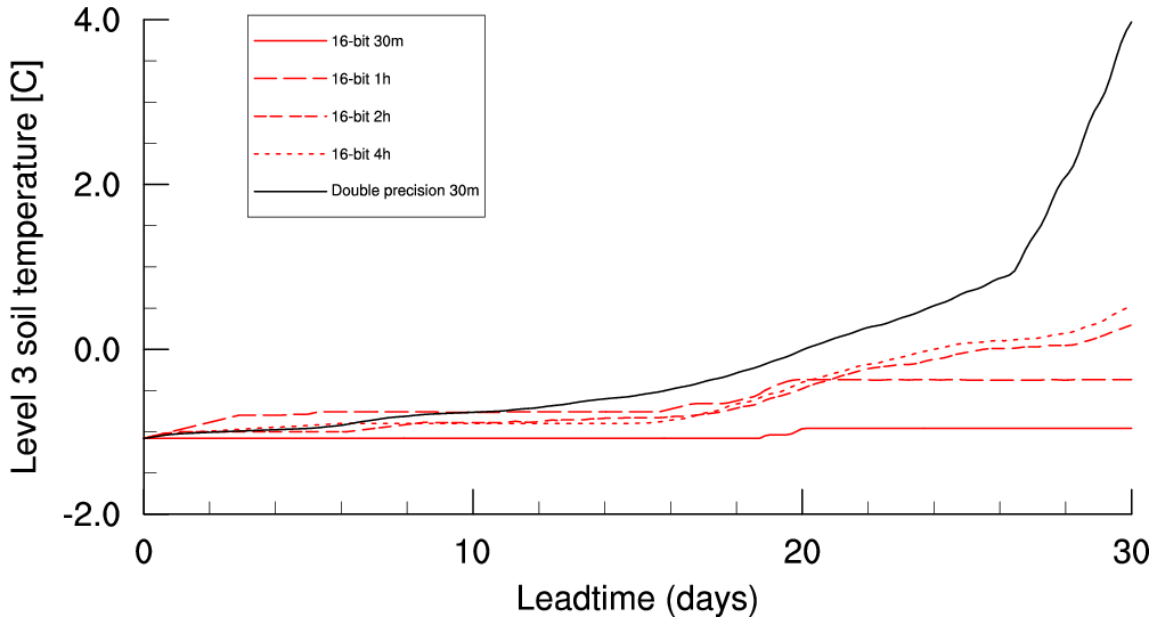
$$\frac{\partial T}{\partial t} = D\frac{\partial^2 T}{\partial z^2}$$

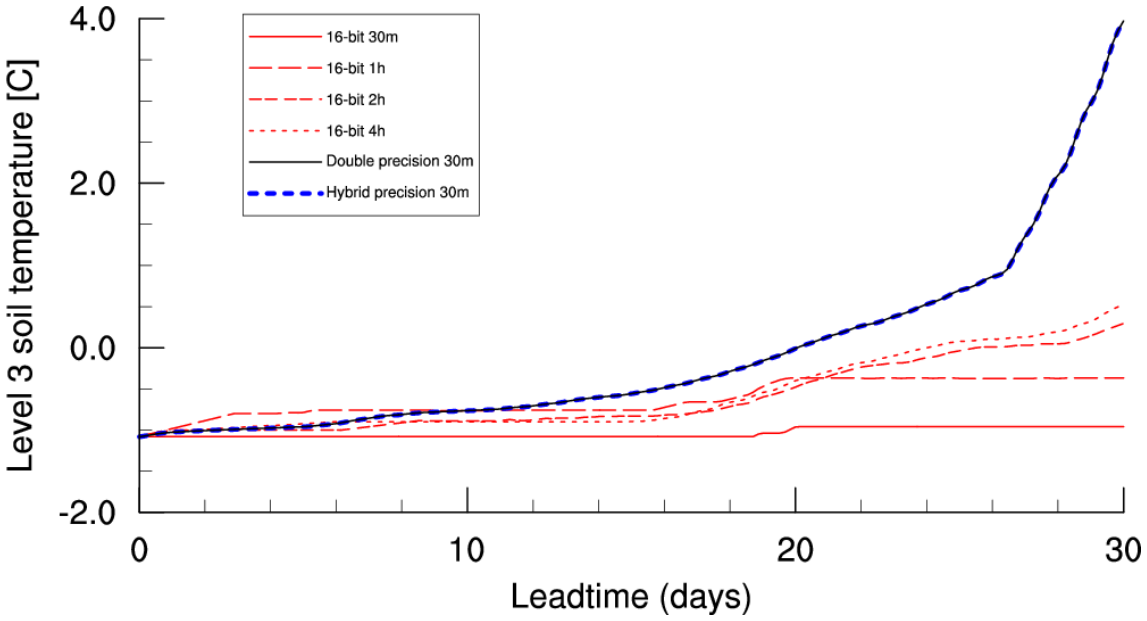
$$T_j^{n+1} = T_j^n + \Delta t\, D\, \frac{(T_{j+1}^n - 2T_j^n + T_{j-1}^n)}{(\Delta z)^2}$$

64 bits

32 bits





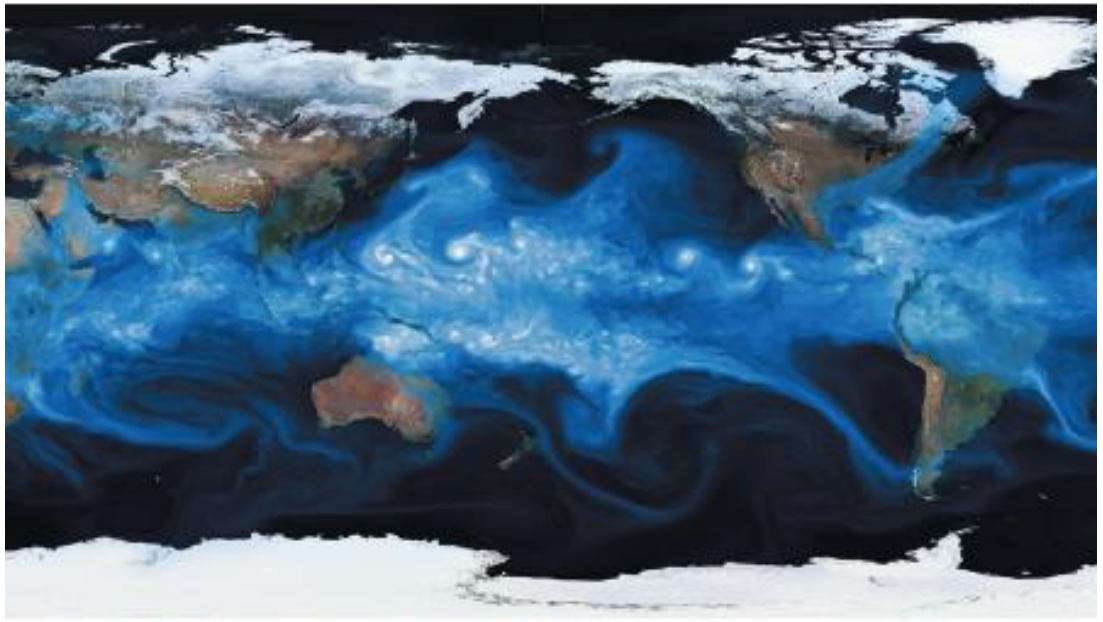

Dave Macleod,
Peter Düben,
Andrew Dawson

# Less Precision, More Accuracy

- Consider an attempt to simulate some complex multi-dimensional nonlinear system
- Some computations (X) must be computed precisely
- Some computations (Y) are inherently uncertain and therefore do not need to be computed precisely.
- Simulating the system is constrained by available energy, such that only a small subset Y(precise) of Y can be computed precisely.
- Imprecise computations are energetically less demanding and allow a larger subset Y(imprecise) of Y to be computed imprecisely
- Is the overall accuracy of the system represented by X(precise)+Y(imprecise) greater than that represented by X(precise)+Y(imprecise)?

# Possible Applications

- Weather/Climate
- Plasmas in tokamaks
- Astrophysics
- Cosmology
- Turbulent flow
- Combustion
- The Brain
  - Simulating the brain
  - The actual brain (energy limited to c. 50W)



A simulation of Earth's atmosphere generated by the Community Atmosphere Model.

# Build imprecise supercomputers

Energy-optimized hybrid computers with a range of processor accuracies will advance modelling in fields from climate change to neuroscience, says **Tim Palmer**.

Today's supercomputers lack the power to model accurately many aspects of the real world, from the impact of cloud systems on Earth's climate to the processing ability of the human brain. Rather than wait decades for sufficiently powerful supercomputers — with their potentially unsustainable energy demands — it is time for researchers to reconsider the basic concept of the computer. We must move beyond the idea of a computer as a fast but otherwise traditional "Turing machine", churning through calculations bit by bit in a sequential, precise and reproducible manner.

In particular, we should question whether all scientific computations need to be performed deterministically — that is, always producing the same output given the same input — and with the same high level of precision. I argue that for many applications they do not.

Energy-efficient hybrid supercomputers with a range of processor accuracies need to be developed. These would combine conventional energy-intensive processors with low-energy, non-deterministic processors, able to analyse data at variable levels of precision. The demand for such machines could be substantial, across diverse sectors of the scientific community.

## MORE WITH LESS

Take climate change, for example. Estimates of Earth's future climate are based on solving nonlinear (partial differential) equations for fluid flow in the atmosphere and oceans. Current climate simulators — typically with grid cells of 100 kilometres in width — can resolve the large, low-pressure weather systems typical of mid-latitudes, but not individual clouds. Yet modelling cloud systems accurately is crucial for reliable estimates of the impact of anthropogenic emissions on global temperature[1].

The resolution of this computational grid is determined by the available computing power. Current petaflop computers can perform up to $10^{15}$ additions or multiplications — floating-point operations — per second (flops). By the early 2020s, next-generation exaflop supercomputers, capable of $10^{18}$ operations per second, will be able to resolve the largest and most vigorous types of thunderstorm[2]. But cloud physics on scales smaller than a grid cell will still have to be approximated, or