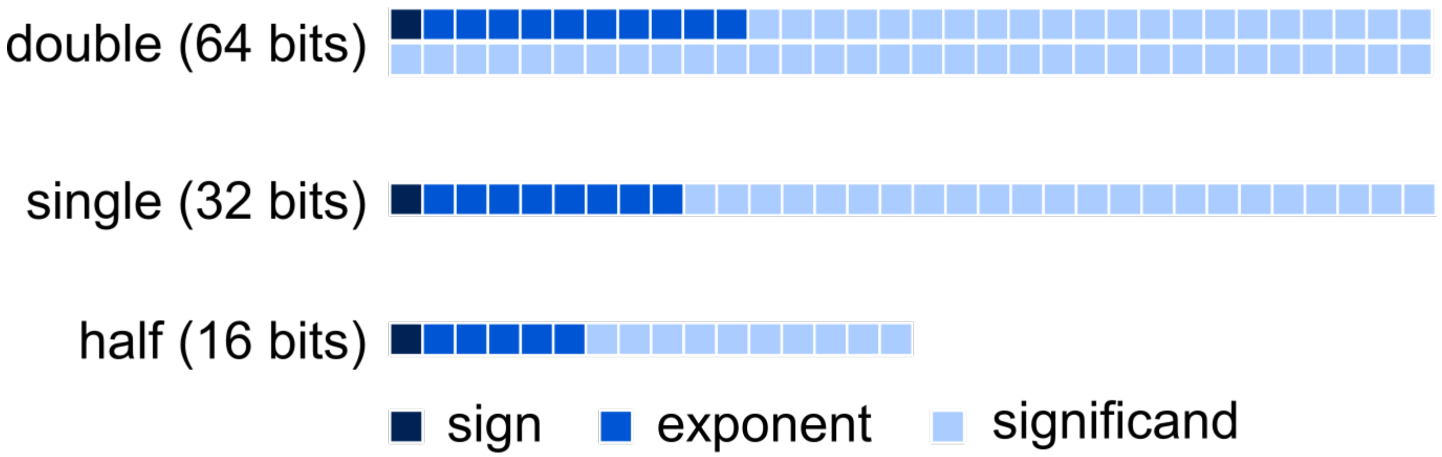
Reducing Precision in Ensemble Data Assimilation

Sam Hatfield, Peter Düben, Matthew Chantry, Tim Palmer
(also Aneesh Subramanian, Keiichi Kondo, Takemasa Miyoshi)
samuel.hatfield@physics.ox.ac.uk

Floating point arithmetic



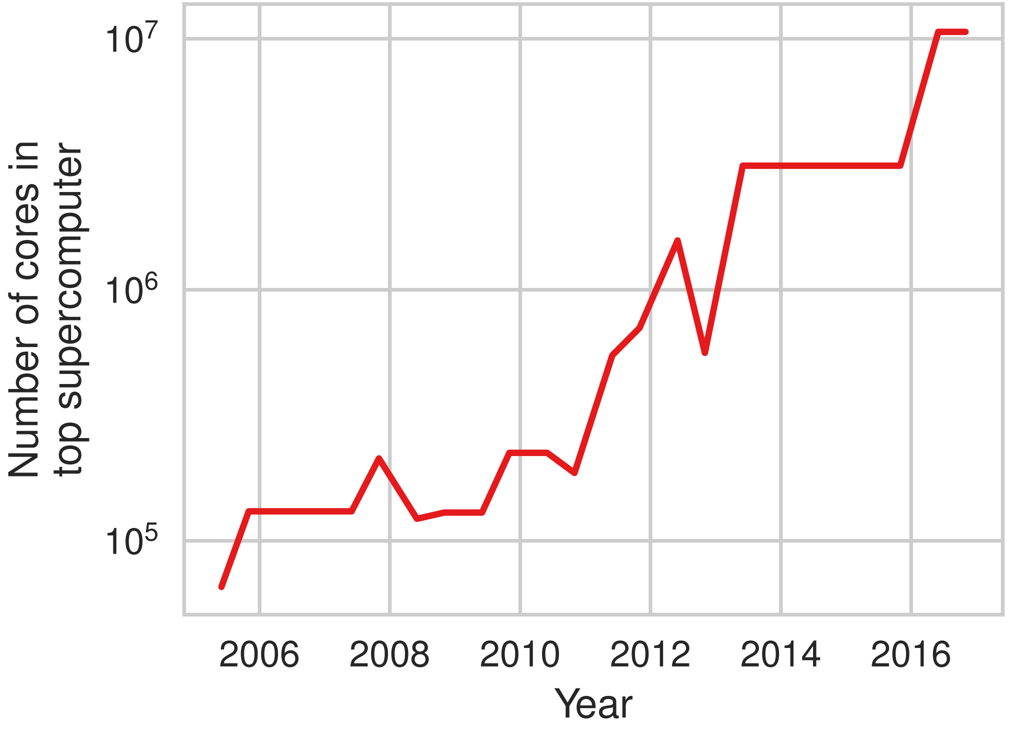
lower precision

Example number:

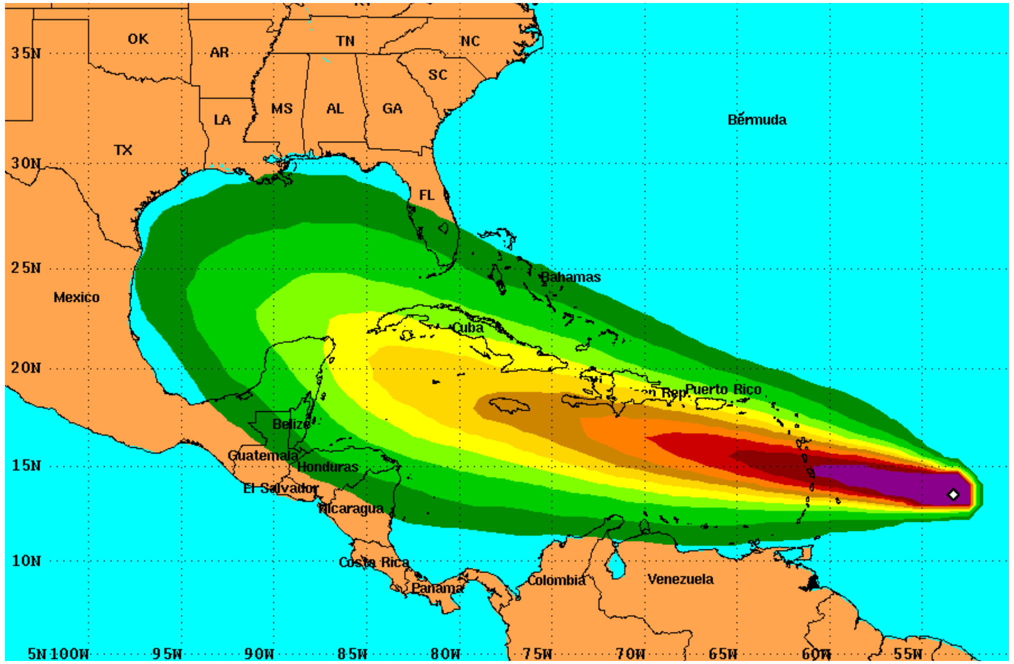
$$-4.938 = \underbrace{-1}_{\text{sign}} \times (1+\underbrace{0.2345}_{\text{significand}}) \times 2^{\overbrace{2}^{\text{exponent}}}$$

Why reduce numerical precision?

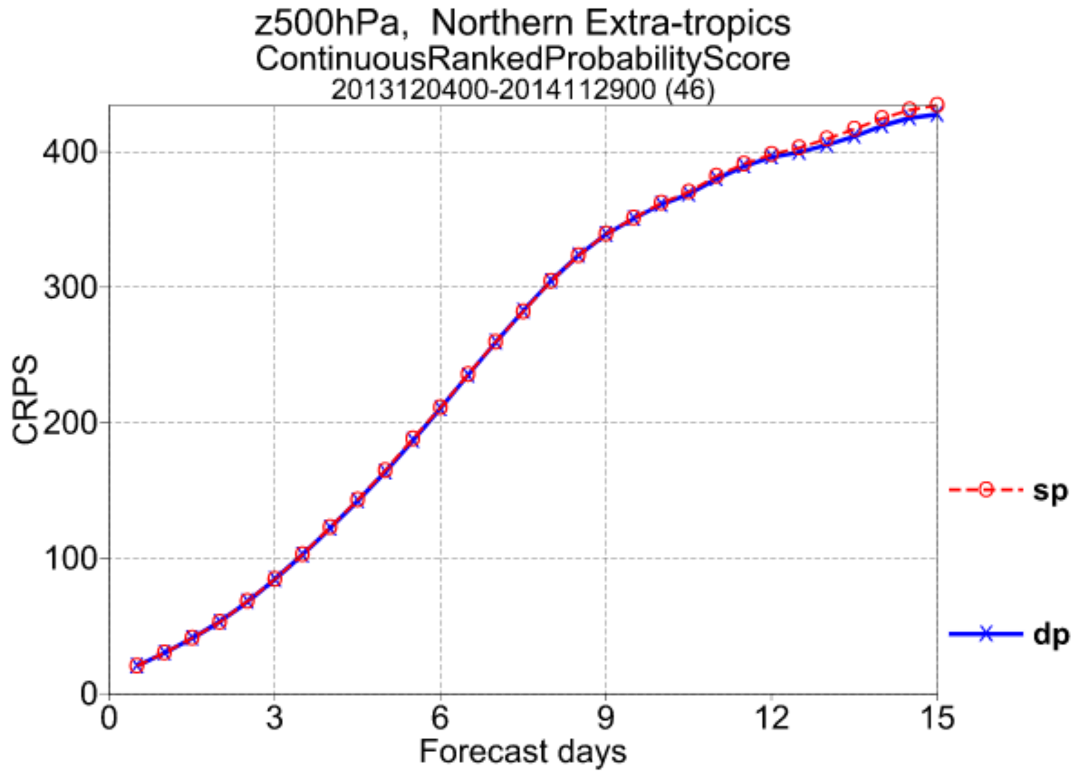
1) Computing trends



2) Model uncertainties

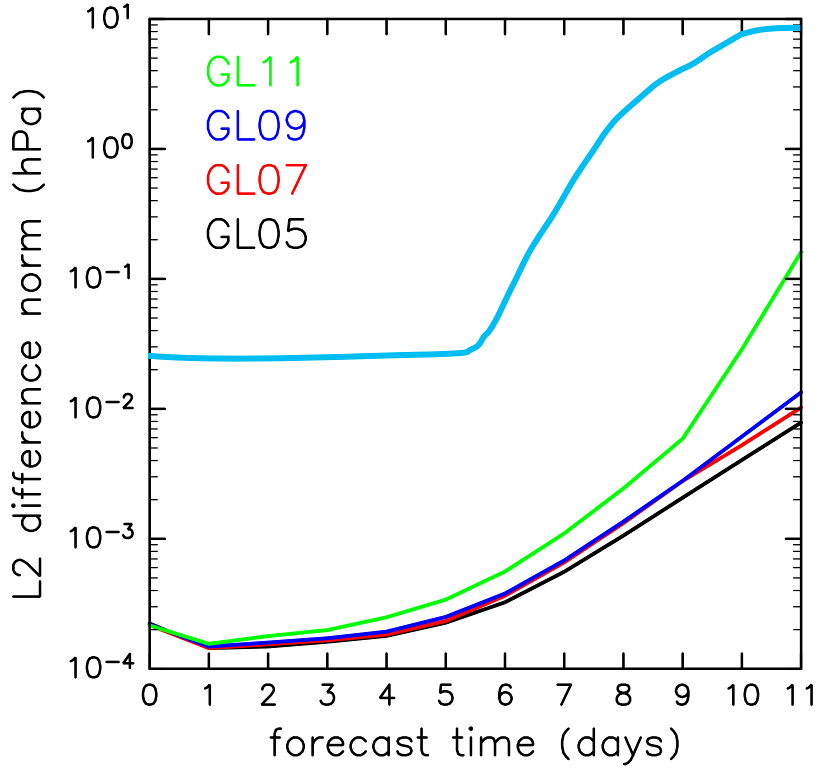


Single precision in NWP



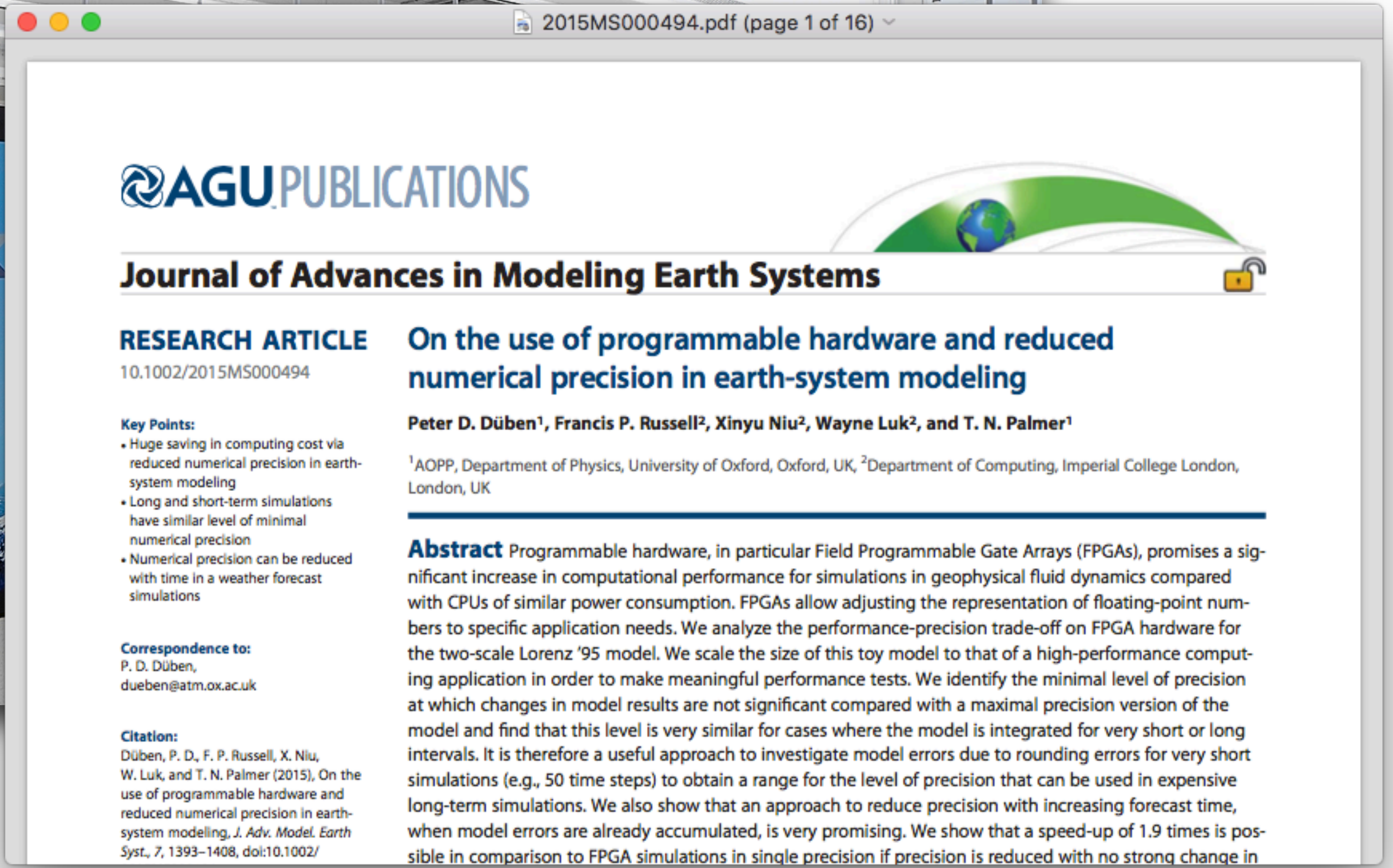
ECMWF's Integrated Forecasting System
(Váňa et al., 2016, MWR)

Jablonowski-Williamson benchmark (QJRMS 2006)



NICAM (non-hydrostatic, icosahedral)
(Nakano et al., 2018, MWR)

Below single precision



AGU PUBLICATIONS

Journal of Advances in Modeling Earth Systems

RESEARCH ARTICLE

10.1002/2015MS000494

Key Points:
- Huge saving in computing cost via reduced numerical precision in earth-system modeling
- Long and short-term simulations have similar level of minimal numerical precision
- Numerical precision can be reduced with time in a weather forecast simulations

Correspondence to:
P. D. Düben,
dueben@atm.ox.ac.uk

Citation:
Düben, P. D., F. P. Russell, X. Niu, W. Luk, and T. N. Palmer (2015), On the use of programmable hardware and reduced numerical precision in earth-system modeling, *J. Adv. Model. Earth Syst.*, *7*, 1393–1408, doi:10.1002/

On the use of programmable hardware and reduced numerical precision in earth-system modeling

Peter D. Düben[1], Francis P. Russell[2], Xinyu Niu[2], Wayne Luk[2], and T. N. Palmer[1]

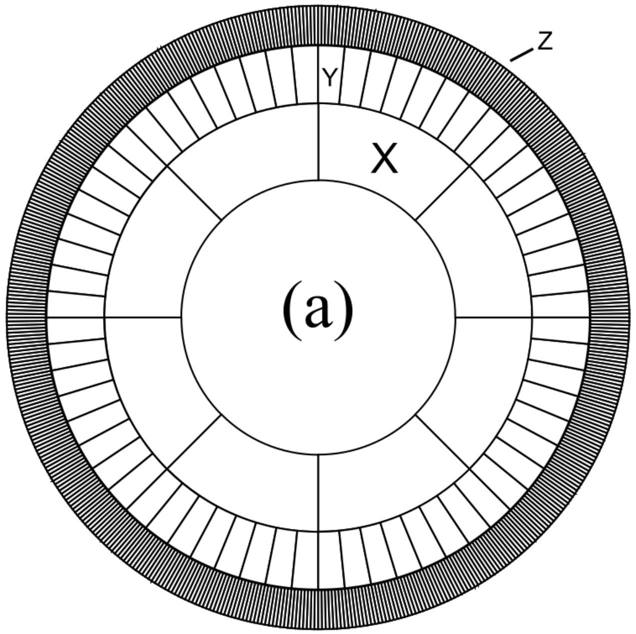
[1]AOPP, Department of Physics, University of Oxford, Oxford, UK, [2]Department of Computing, Imperial College London, London, UK

Abstract Programmable hardware, in particular Field Programmable Gate Arrays (FPGAs), promises a significant increase in computational performance for simulations in geophysical fluid dynamics compared with CPUs of similar power consumption. FPGAs allow adjusting the representation of floating-point numbers to specific application needs. We analyze the performance-precision trade-off on FPGA hardware for the two-scale Lorenz '95 model. We scale the size of this toy model to that of a high-performance computing application in order to make meaningful performance tests. We identify the minimal level of precision at which changes in model results are not significant compared with a maximal precision version of the model and find that this level is very similar for cases where the model is integrated for very short or long intervals. It is therefore a useful approach to investigate model errors due to rounding errors for very short simulations (e.g., 50 time steps) to obtain a range for the level of precision that can be used in expensive long-term simulations. We also show that an approach to reduce precision with increasing forecast time, when model errors are already accumulated, is very promising. We show that a speed-up of 1.9 times is possible in comparison to FPGA simulations in single precision if precision is reduced with no strong change in

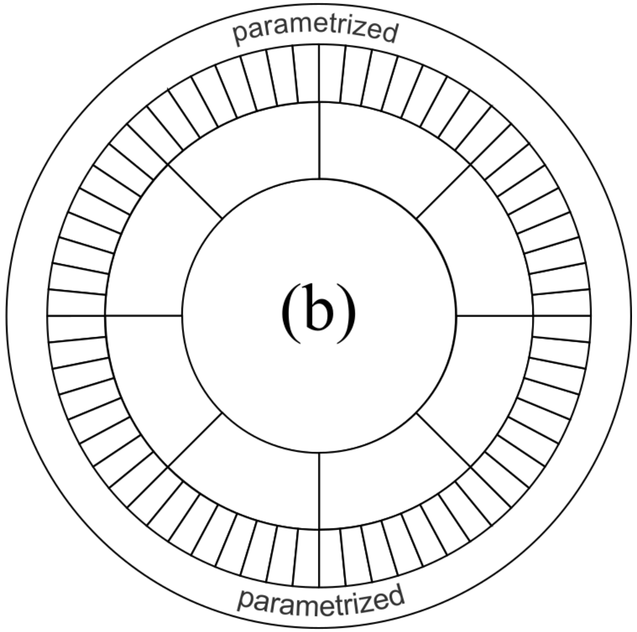
Reducing precision in data assimilation

- Research questions:
    - Can we adjust precision to a level justified by “system uncertainty”? (e.g. model error, observation error)
    - Can we improve the quality of analyses if we reinvest computational savings from reducing precision? (e.g. boost the ensemble size)
- Talk outline:
    1. Lorenz ‘96/ensemble square root filter
    2. SPEEDY/local ensemble transform Kalman filter

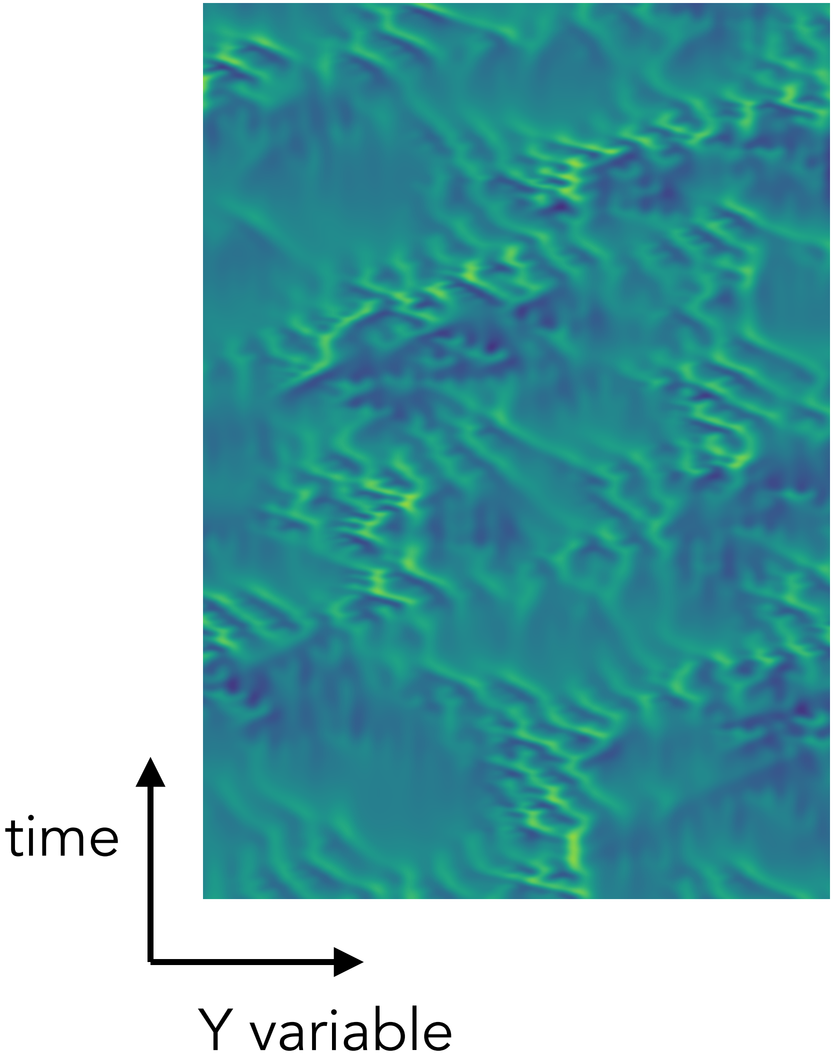
Lorenz '96



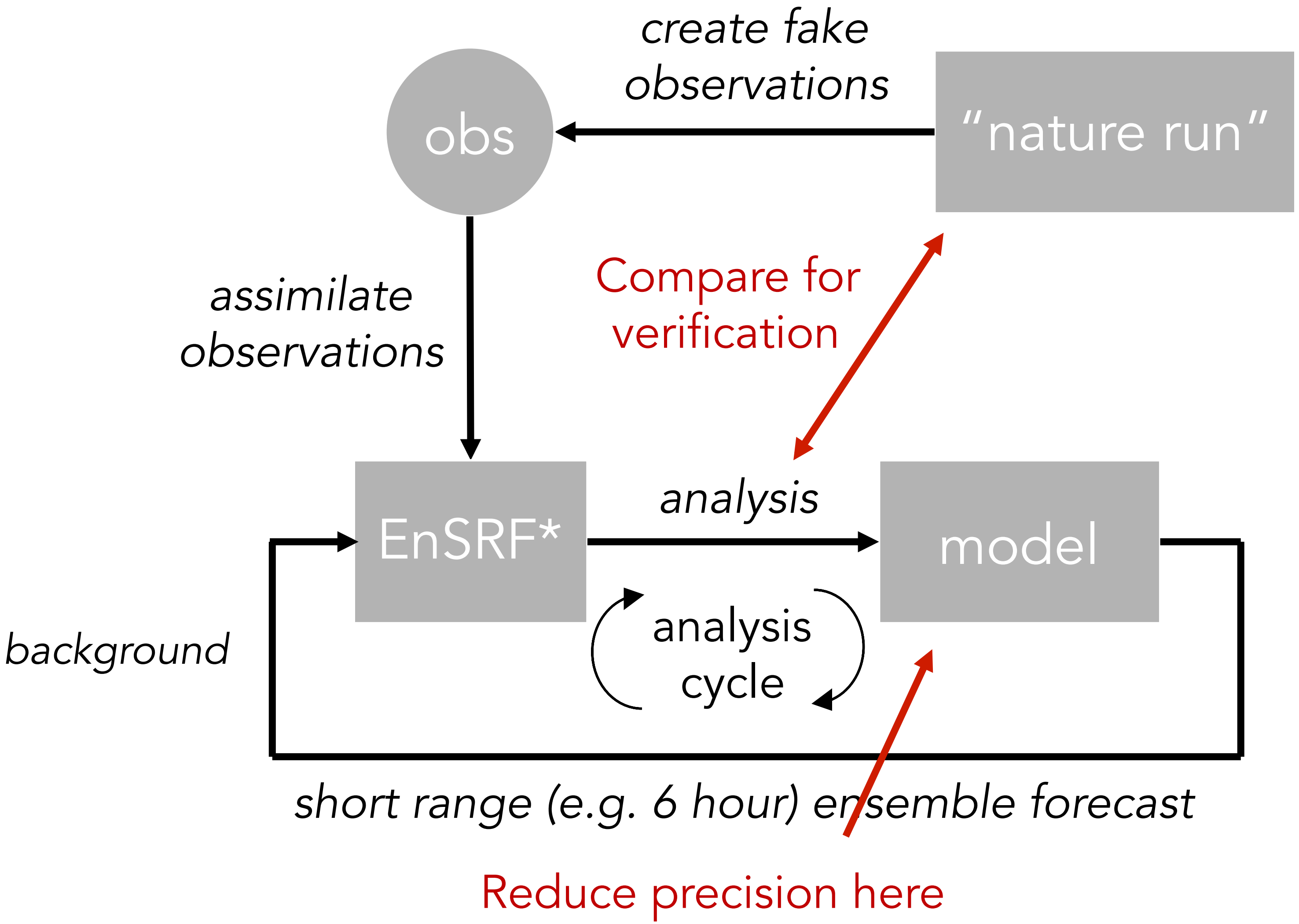
Nature model
(used for synthetic observations)



Model used for
data assimilation
(simplified)



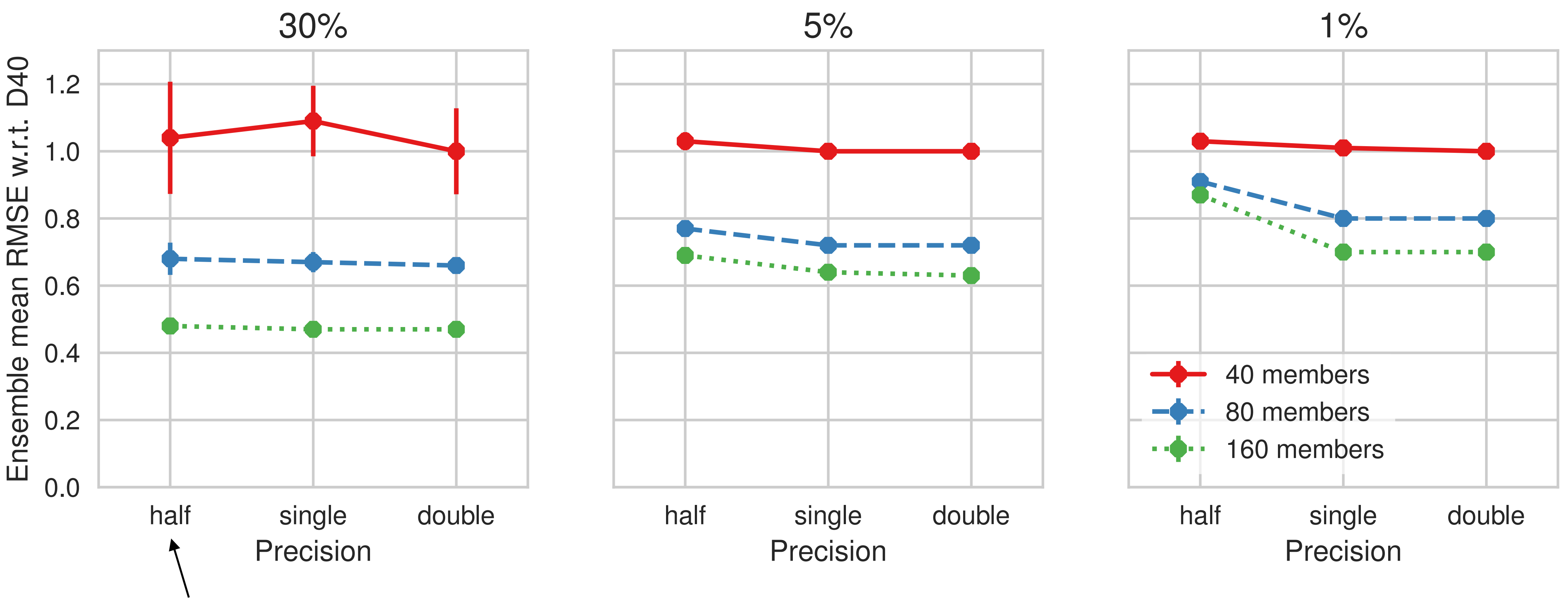
Assimilation setup



*ensemble square root filter

Reduced precision analyses

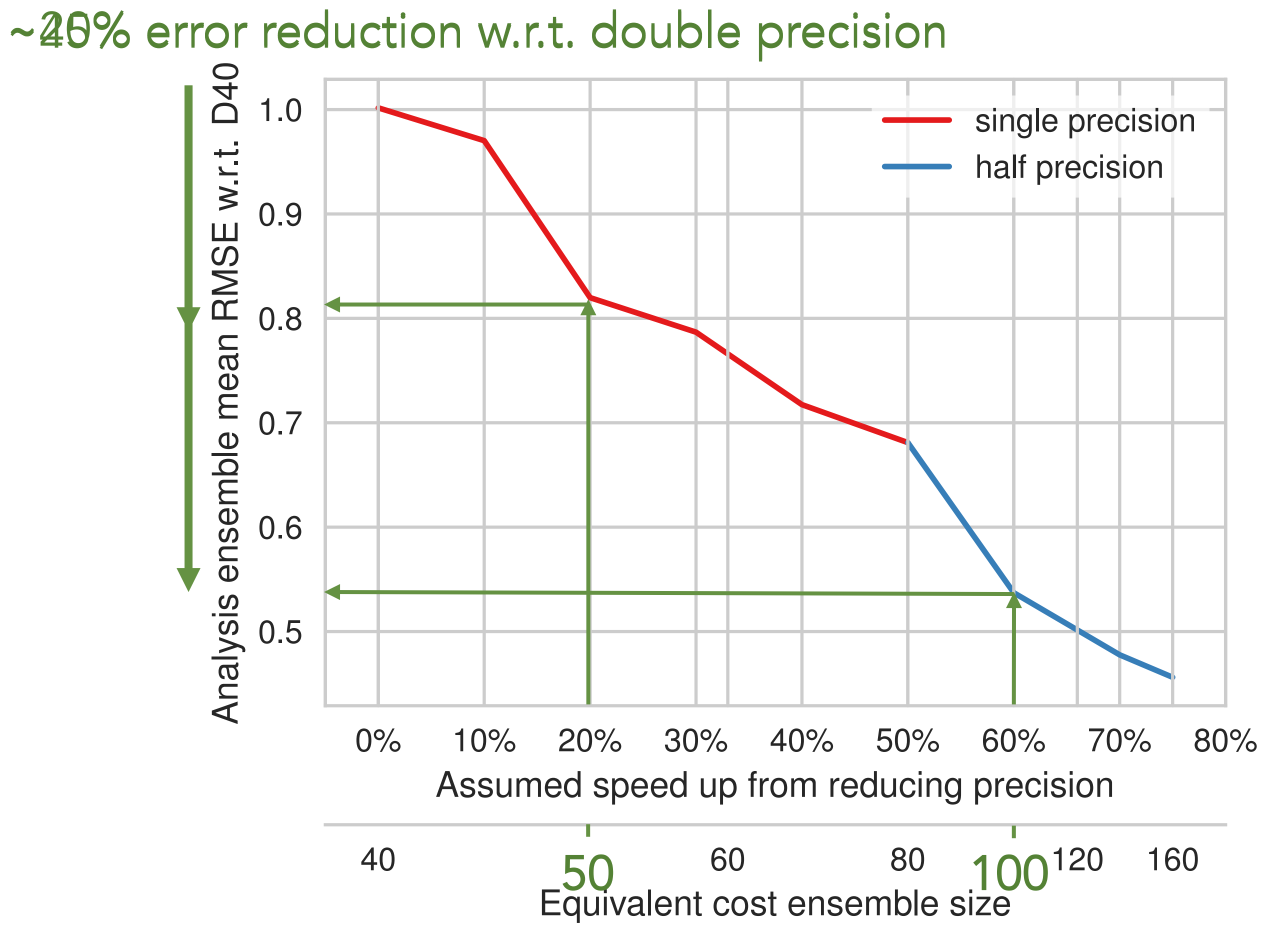
Observation error (% of natural variability)



half-precision model used a software emulator

Precision/ensemble size trade-off

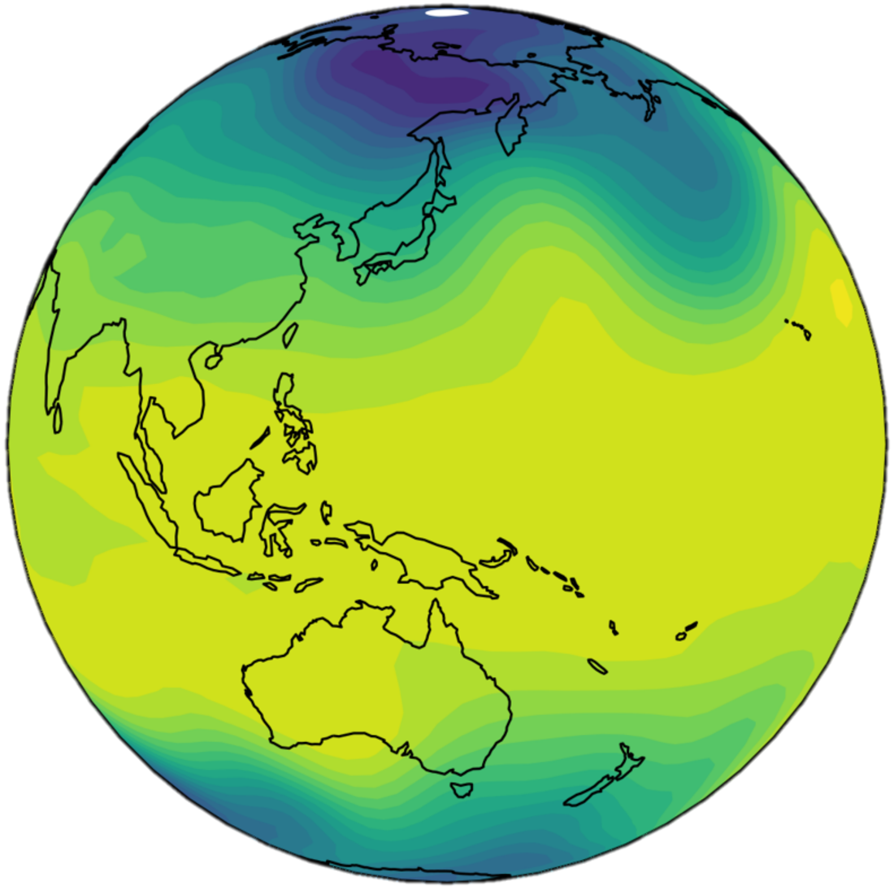
- Trade bits for ensemble members
- Computational speed-up from reducing precision is difficult to predict
- But even a modest speed-up can be beneficial



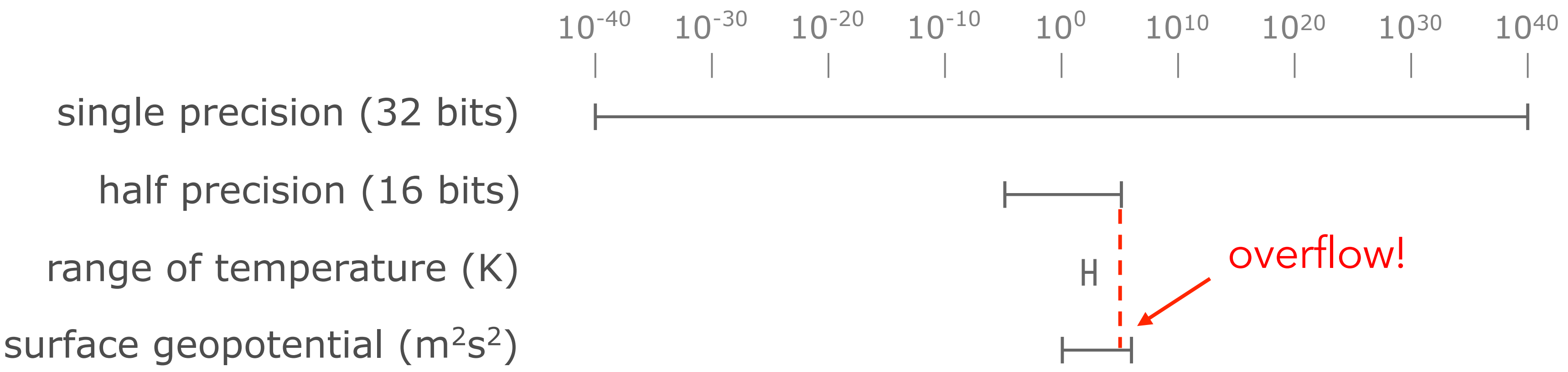
Hatfield et al. MWR 2018

SPEEDY

- Spectral dynamical core
- T30 resolution – roughly 400km at equator
- Several parametrized processes: convection, radiation, land/sea fluxes
- Reduce precision in forecast model → measure change in analysis error



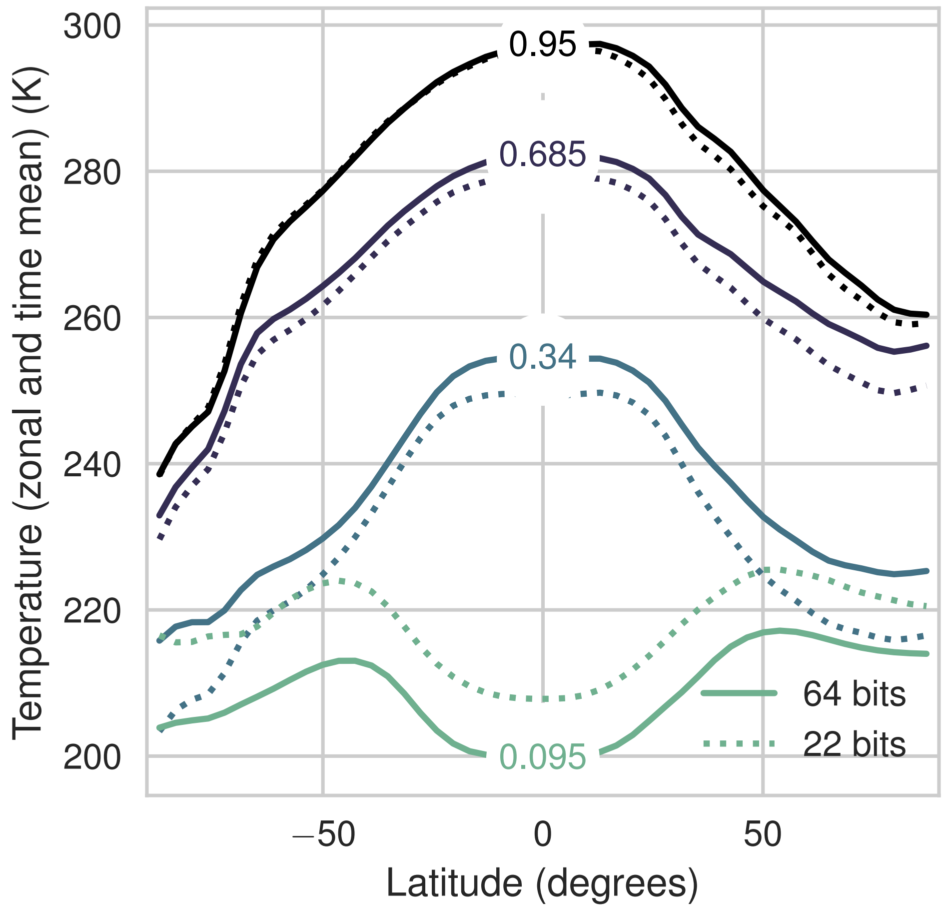
Using half precision arithmetic in SPEEDY

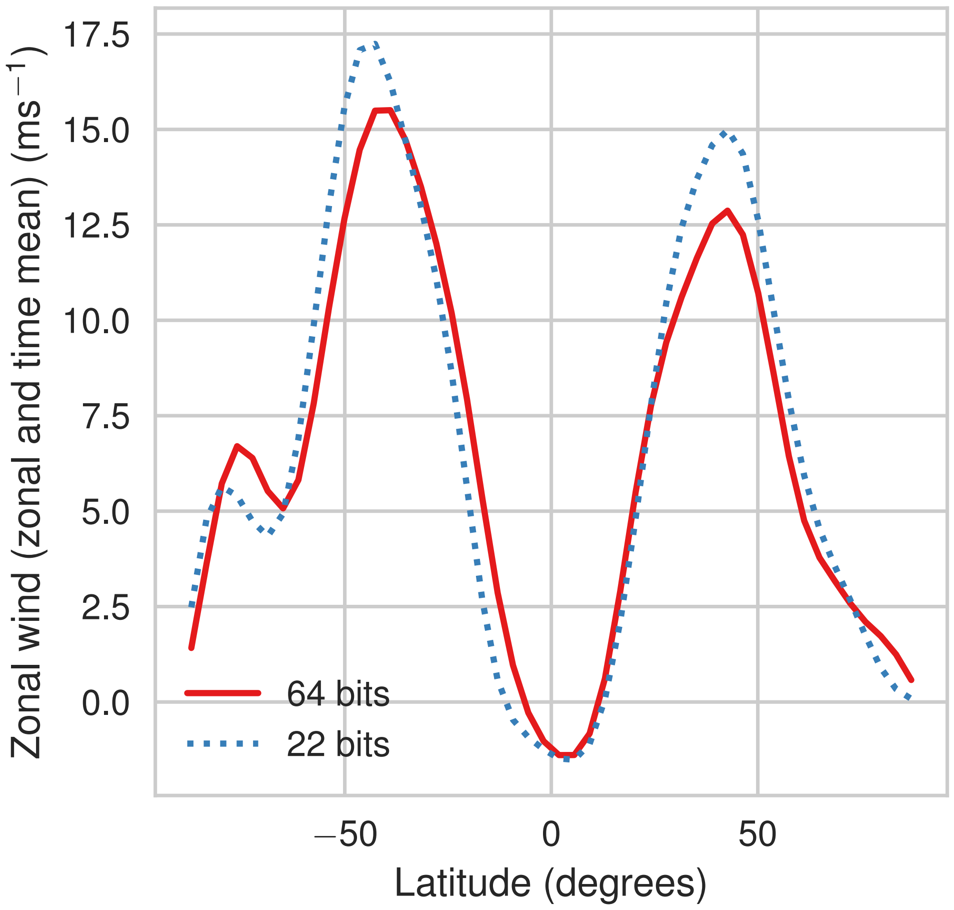


- Half-precision floats have a limited range (10^{-5} ~ 10^{6})
- For now, only reduce significand width (52 bits → 10 bits)
- Compare 22 bits (1+11+10) with 64 bits (1+11+52)

sign
exponent
significand

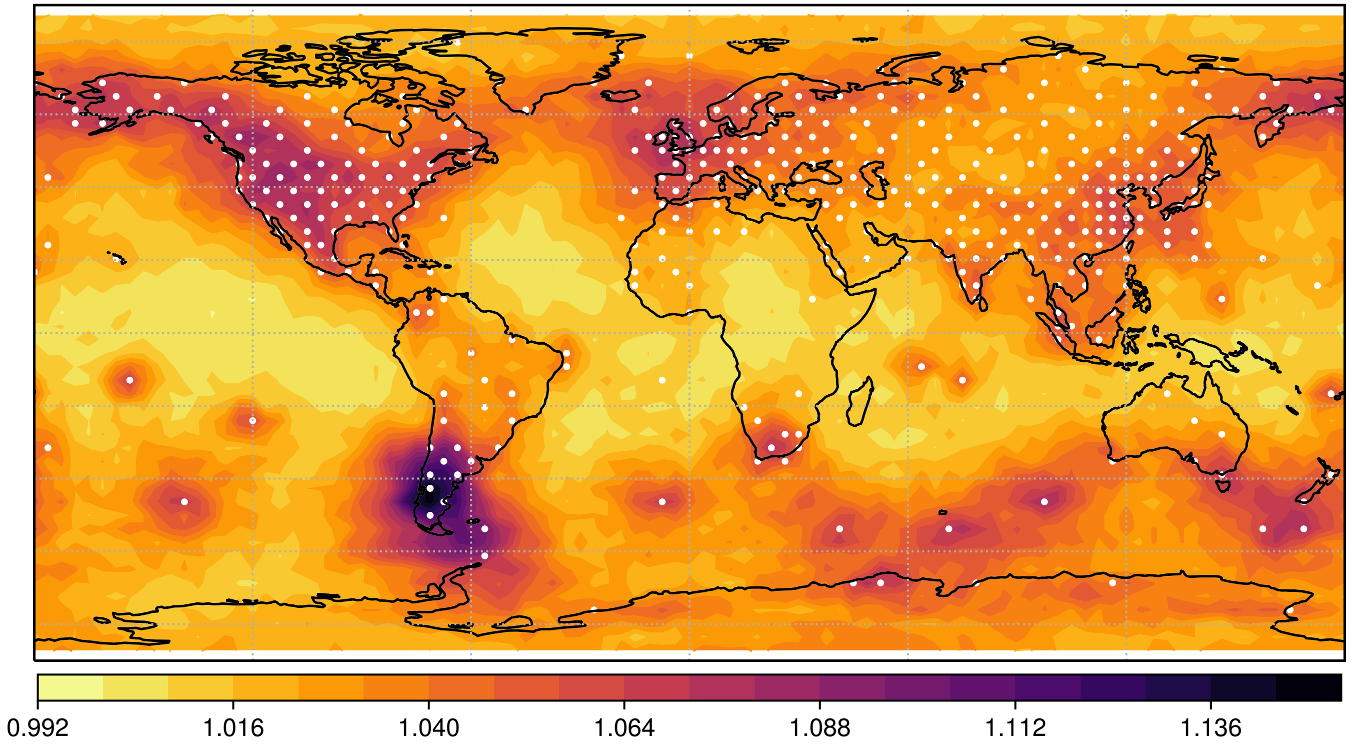
Reduced precision SPEEDY biases





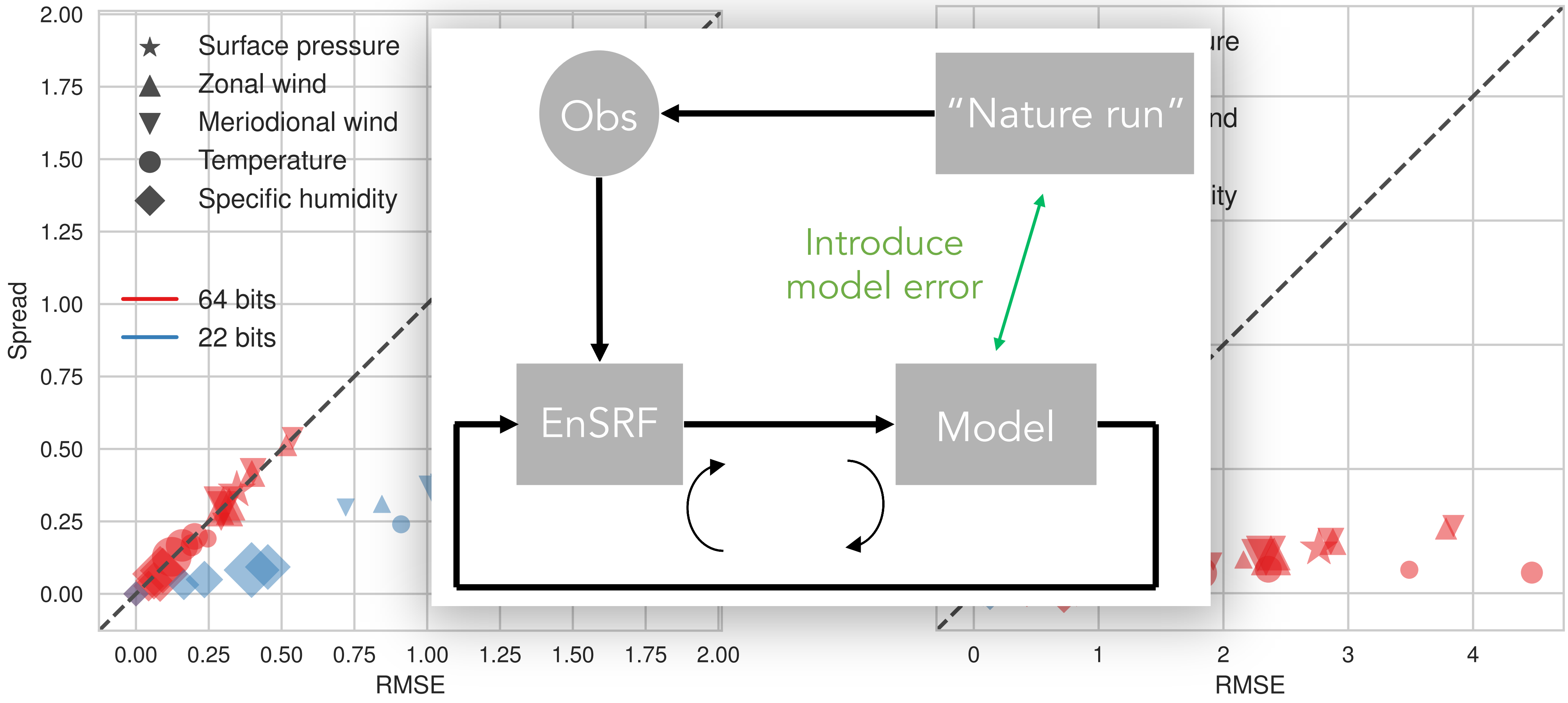
Assimilation setup

- Assimilation algorithm: local ensemble transform Kalman filter
- Synthetic observations
- 20 members
- Gaspari-Cohn covariance localisation and RTPP inflation



Contours: RTPP inflation factor
Dots: observation locations

Perfect model experiments

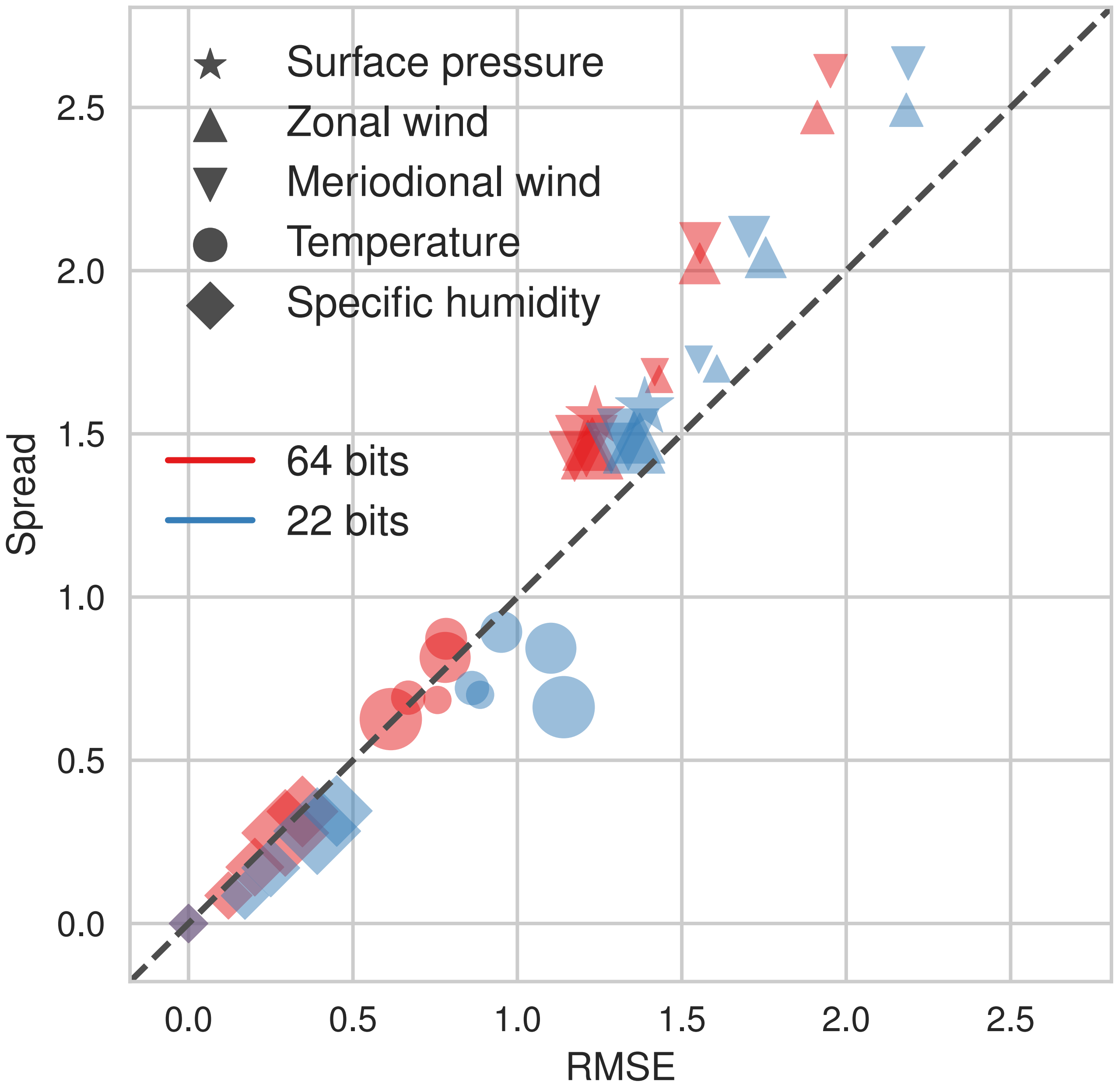


64 bit nature run

22 bit nature run

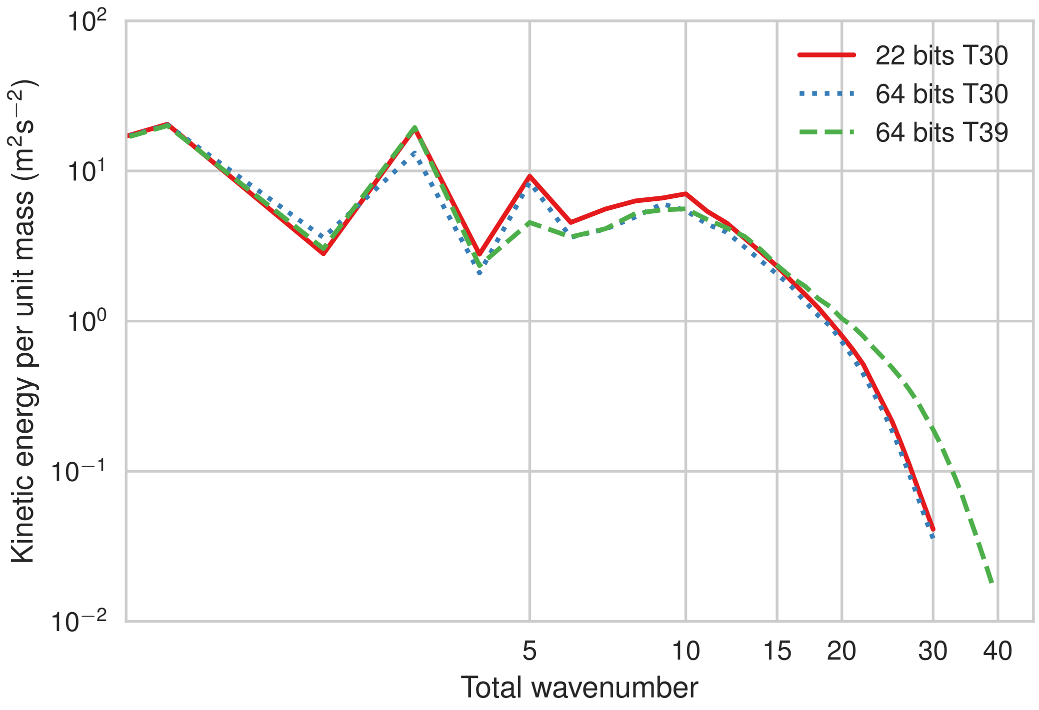
Introducing model error (1)

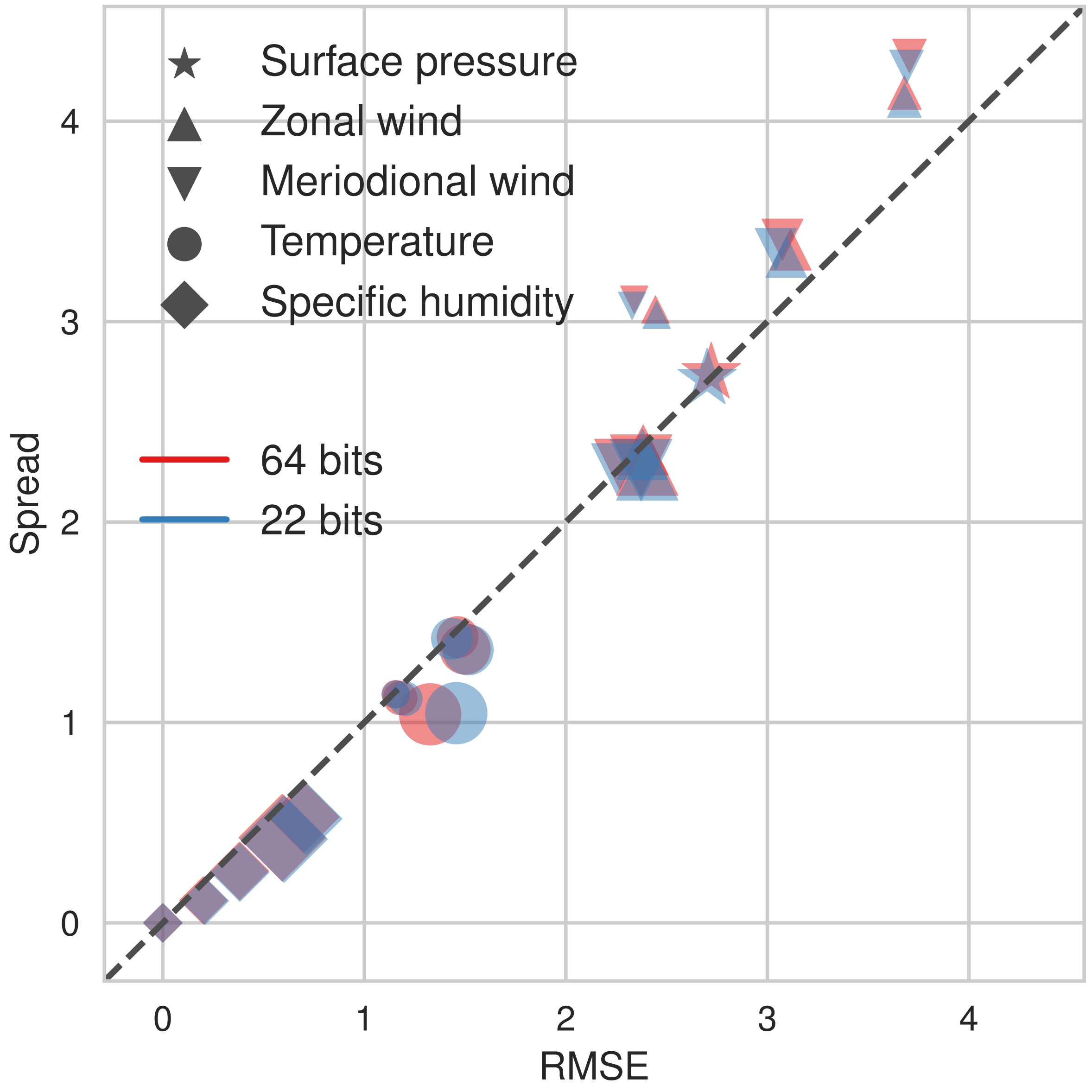
- Try doubling diffusion time scale in assimilation model



Introducing model error (2)

- What about a higher resolution nature run? (T39 instead of T30)





Conclusion

- Reducing precision could provide a one-off ”boost” of computer resources – on the order of a computer upgrade
- The lowest possible precision is constrained by the level of uncertainty (observations, model error etc.)
- The use of half-precision in data assimilation/modeling remains an open question due to the severely limited range

Time to reconsider IEEE floats?

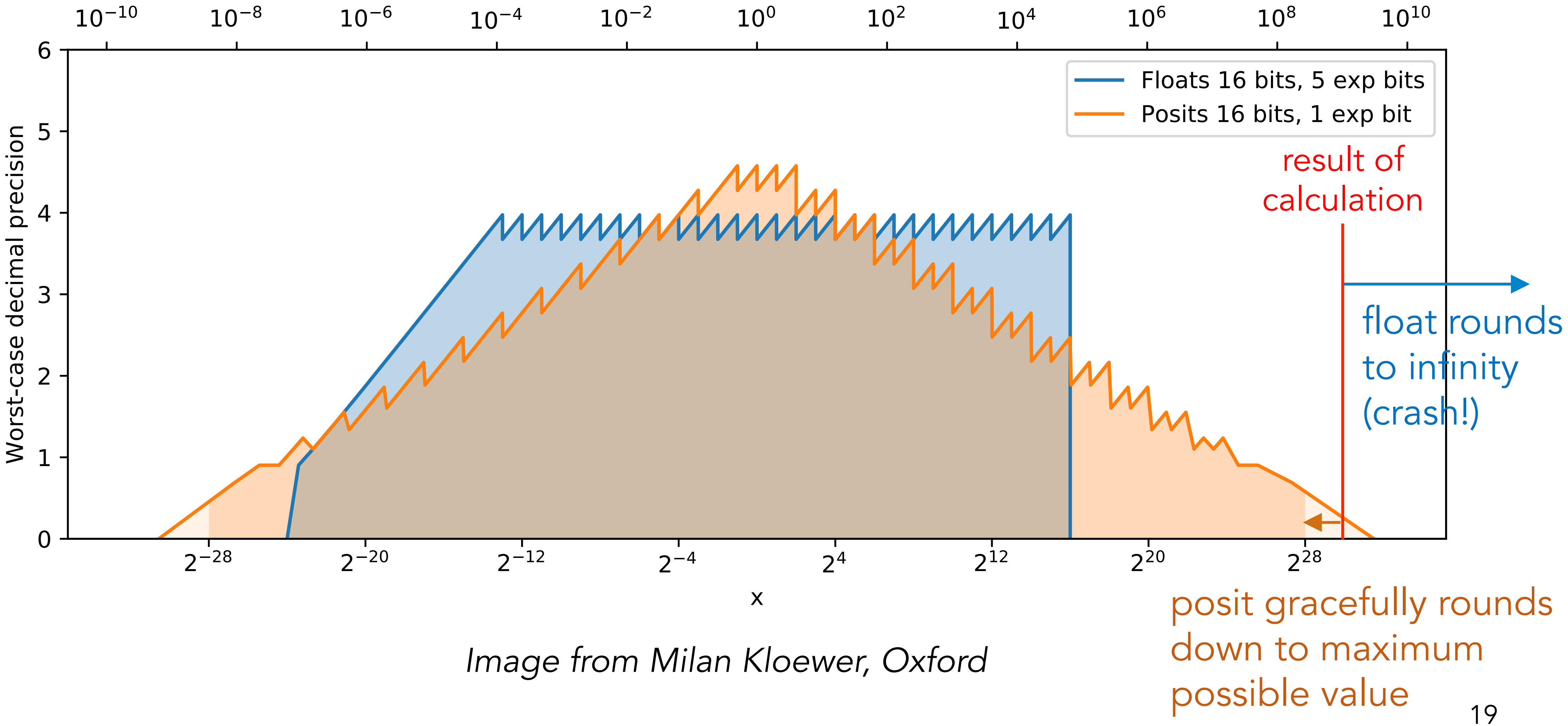


Image from Milan Kloewer, Oxford