Can Reducing Precision Improve Accuracy in Weather and Climate Models?

Tobias Thornes

Abstract

Better weather and climate forecasts are needed to maximise the ability of societies worldwide to prepare for climate change. This paper summarises work regarding the hypothesis that increasing the resolution of global models to achieve this could be done with existing computer technology by reducing the precision of the model variables to increase efficiency. A method of testing this idea carried out using computer simulations of the idealised Lorenz '96 atmosphere is explained. Preliminary results suggest that high-resolution, reduced precision forecasts outperform low-resolution ones. If these results also apply to real-world models, the technique could yield substantial benefits.

Key Words

Inexact hardware Reduced precision Modelling Computing Lorenz Idealised models Forecasting

1. Introduction

The need for accurate forecasts of weather and climate across the whole planet has never been more pressing. Climate change poses a double-headed threat to human well-being, bearing with it the prospect of more frequent extreme weather events that directly endanger lives and infrastructure as well as considerable long-term climatic changes that may render entire ways of life untenable in some regions. If societies are to adapt in time to mitigate these effects and minimise the suffering they will cause, it is imperative that weather and climate forecasts are as accurate as possible. However, there remain considerable uncertainties surrounding, for example, the impact of the hydrological cycle on climate change (Palmer, 2014a) and weather forecasts accurate enough to predict extreme weather events many days or weeks in advance are lacking.

Weather and climate forecasts are heavily dependent upon numerical models of the global atmosphere run on supercomputers. A key limitation on the accuracy of these models comes from the fact that they cannot resolve small-scale features such as convective clouds, which are less than 1 km across (Palmer, 2014b). Currently, the highest-resolution global weather models have a grid-point spacing of 10 km or more, as illustrated in Figure 1, which means that they can only resolve features of the order of 50 km across or larger. The maximum resolution is in turn limited by the speed and memory constraints of today's supercomputing centres; each halving of the grid-point spacing of a model typically also requires a doubling of the temporal resolution, and around an eight-fold increase in computer power is necessary to achieve this (a power of two for each of the one temporal and two spatial dimensions).

Since the dawn of Numerical Weather Prediction (NWP) in the 1960s, forecasters have benefited from the fact that computers have become faster and faster to achieve better and better forecasts. But engineers are now reaching practical limits on the speed of individual computer processors: for example, as the density of computer transistors increases they become increasingly difficult to keep cool enough to function (Markov, 2014). This means that many processors have to be run in parallel to maintain the trend towards increasing computer speed, and weather and climate models are having to be extensively re-written to take full advantage of these parallel processors (Wehner, et al., 2008), a lengthy and expensive task. Therefore, using conventional computing techniques it may take many decades before computers have advanced sufficiently to provide forecasts accurate enough to meet the needs of human society.

This has motivated a new approach to computing that aims to improve forecasts much sooner by making computers more efficient through reducing the precision with which numbers are represented during computations (Duben, et al., 2014a). Reducing the precision in itself can only make a numerical model less accurate, but the accuracy may nevertheless improve overall if the resulting computational cost savings can be re-invested to increase the model resolution. Because atmospheric variables such as temperature and pressure are known with less certainty on smaller-scales than larger ones, it has been argued that in double-precision small-scale variables are represented more precisely than is justified by the uncertainty in the observations and the model error (Palmer, 2012). In particular, 'parameterisation' of processes occurring on scales too small to be resolved introduces large uncertainties on the smallest resolved scales. If this is the case, reducing the precision of the smaller-scale variables more than that of larger-scale ones might be expected to have much less of an effect on the accuracy than the corresponding increase in model resolution would have. However, this should only be applied to parts of the model: other parts to which the output is very sensitive – such as the time-stepping scheme – ought to be kept entirely in double-precision to avoid crashes.

The idea behind 'flexible precision', a form of 'inexact' computing, is therefore to maximise the accuracy of a weather and climate model by choosing the optimal level of precision for each number used in each part of a model. For a given computation, there will exist an optimal balance between precision and resolution, depending on how small a scale the variables it involves can be measured at and the uncertainty in the measurement. The hypothesis is that by reducing the precision in some of the variables and using the memory and processing time that this saves to increase the model resolution instead, these flexible-precision forecasts can be made to be more accurate than conventional forecasts. This would require new hardware capable of carrying out different computations in the model at different levels of precision, which, although not yet available in a form suitable for use in supercomputer centres, is already under development, with half-precision processors soon to be available (see Section 2). Figure 2 compares this new approach with the conventional, double-precision computing approach diagrammatically.

This paper is intended to provide an introduction to the field of reduced precision computing in the context of weather and climate models and to provide a first taste of new work being carried out by the author to investigate the effects of the 'inexact' technique. This work will add to a number of recent studies demonstrating the potential advantages of efficient, 'inexact' computing for forecasts using models of both hypothetical, 'idealised' atmospheres (Duben, et al., 2014) and the real Earth atmosphere (Duben & Palmer, 2014). Section 2 explores the principles behind the 'inexact' hardware which is being investigated by such researchers as a means of increasing the accuracy of models by reducing the precision at which they are run. Section 3 outlines the methodology behind tests of the 'flexible precision' inexact method in the Lorenz '96 system, an idealised atmospheric model. The implications for global forecasts of the preliminary results obtained during such tests, which will be published in full elsewhere, are summarised in Section 4.

2. 'Inexact' Hardware for Weather and Climate Simulators

Any computer has a finite number of 'bits' of information that it can process in a given time. The speed at which it can run a program depends on the number of computer 'bits' used to represent each number in the calculations that the program includes. Conventionally, computers operate in 'double precision', where sixty-four bits are used for each number, which usually ensures that rounding errors in the calculations are negligible. These can be carried out much faster in single- or half- precision ('reduced precision'), which use fewer bits (see Table 1), but because the numbers are less precise the calculations are less likely to be exactly correct.

A number x is represented on a computer in binary form, which means that the bits allocated to that number are split between three parts: the 'sign' ± 1 (whether the number is positive or negative); the 'exponent' E, which is the next lowest power of two to the number; and the 'significand' S, which is the fraction, always between 1 and 2, by which this power must be multiplied, as illustrated below:

$$x = \pm 1 \times 2^E \times S.$$

Restricting the number of bits used for the significand restricts the number of decimal places to which the number can be accurately represented. Restricting the number of bits given to the exponent restricts the possible range of values that it can take (see Table 1). If weather and climate models were run in reduced precision one might therefore expect the accuracy to decrease, unless this is compensated for by using the savings in computational cost to improve the model in some other way.

Flexible precision computing involves carrying out different calculations at different levels of precision. To be run in this way, models will not necessarily need to be extensively re-written, but they will need to use so-called 'inexact' hardware where the precision used for each computation can be chosen explicitly. Such hardware doesn't yet exist in a form suitable for running global weather and climate models in flexible precision, but could be built by combining existing double-, single- and half-precision technology into a single processor if manufacturers and weather centres are persuaded that flexible precision will bring real benefits. Already, positive developments in this direction are being made. Prototype 'probabilistic pruning' chips have been made that can de-activate parts of a circuit that are deemed less significant than others to save energy (Palem & Lingamneni, 2013), and a similar approach might be used to re-route some computations into single- or half-precision parts of a processor instead. In 2016, NVIDIA will release its 'Pascal' Graphical Processing Unit (GPU), the company's first high-performance GPU capable of operating in double-, single- and half-precision, in recognition of the need for better energy-efficiency to increase computer speed in applications ranging from high-performance gaming to low-energy handheld devices (Moammer, 2015). In combination, these components might be scaled up to deliver flexible-precision supercomputers in forecast centres within a few years.

In order to bring this about, it is necessary to demonstrate that producing flexible-precision supercomputer hardware and using it for weather and climate forecasts would bring worthwhile benefits. However, in the absence of flexible-precision hardware it is impossible to explicitly test its effects on forecast accuracy without emulating the hardware on conventional computers. In this study, such emulation is done by rounding the numbers used in some of the calculations as though they were being represented in reduced precision whilst leaving the double-precision calculations unchanged. Whenever an operation such as addition or multiplication is carried out, the input numbers are passed to the emulator, which carries out the operation then rounds the output number appropriately before passing it back to the main program. This enables the effects of reduced precision to be tested using conventional hardware.

Applying such an emulator is computationally costly, so this method cannot be used to quantify the computational cost savings associated with flexible precision. But it can be used to investigate whether high-resolution forecasts made with some of the variables in reduced precision match or exceed the accuracy of low-resolution forecasts made with all variables in double-precision, assuming that these would incur similar computational costs.

3. The Lorenz '96 Idealised Model Atmosphere: A Testbed for Reduced Precision

The potential for reduced numerical precision to improve accuracy can be tested in an idealised atmosphere such as that formulated by Lorenz in 1996 (Lorenz, 2006), hereafter referred to as Lorenz

’96. The advantages of using such an idealised atmosphere are two-fold. Because it is much smaller and less complicated than a global model of the real Earth atmosphere, it is possible to perform many experimental model runs in quick succession, unburdened by huge computational costs. Furthermore, because there are so few variables in the model it is possible to define the ‘true’ state of the system, against which all other models can be compared, whereas in the real atmosphere the ‘true’ state is only known as accurately as (limited) observations permit. Meanwhile, the model exhibits chaotic behaviour and non-linear interactions between spatial scales, in this respect resembling the real Earth atmosphere.

For these reasons, Lorenz ’96 has been used extensively in previous studies to test new modelling and data analysis techniques before applying them to models of the real atmosphere, including some initial tests of the effects of ‘inexact’ hardware on computer simulations of the system (Duben, et al., 2014). However, these studies have used Lorenz’ original formulation, in which there are only two scales or ‘tiers’ of variables: slowly-varying, large-scale variables labelled X and fast-varying, moderate-scale variables labelled Y. These are hypothetical atmospheric properties but might correspond, for instance, to density, pressure or velocity on large and small scales in the real atmosphere. The time-evolution of each is governed by a separate differential equation that can be solved numerically on a computer. Where dX_k/dt and dY_{jk}/dt represent the rates of change with time of variables X_k and Y_{jk} respectively, the equations are (Lorenz, 2006),

$$\frac{dX_k}{dt} = X_{k-1}(X_{k+1} - X_{k-2}) - X_k + F - \frac{hc}{b}\sum_{j=1}^{J} Y_{j,k}, \quad (1)$$

$$\frac{dY_{j,k}}{dt} = -cbY_{j+1,k}\left(Y_{j+2,k} - Y_{j-1,k}\right) - cY_{j,k} + \frac{hc}{b}X_k. \quad (2)$$

Here, F is the ‘forcing’ applied to the large-scale variables, h denotes the strength of the interaction between the different scales and b and c are the relative magnitudes and rates of evolution of the X- and Y-variables. The indices j and k are used to label the Y- and X-variables respectively, where there is a total of J smaller-scale Y-variables for each of the K larger-scale X-variables. The real Earth atmosphere is a chaotic system, which means that any errors in the initial conditions used in global climate models will grow exponentially over time, doubling in size about every 1.5 days (Lorenz, 2006). The model parameters have to be chosen to ensure that the idealised model is similarly chaotic to the real atmosphere; $F = 20$, $h = 1$ and $b = c = 10$ has been a popular choice in other studies using the Lorenz ‘96 system (Lorenz, 2006 and Arnold, et al., 2013).

Having two separate scales of variables means that the ‘truth’ can be defined by solving the equations explicitly for both scales, and this can be compared to ‘Low Resolution’ models in which the smaller-scale variables are ‘parameterised’ (approximated) instead of being resolved. Such models resemble models of the real atmosphere, where observational and computational constraints mean that variables below a certain scale inevitably have to be parameterised. But to explicitly see whether reduced-precision can improve the forecast accuracy by increasing the available resolution, which amounts to reducing the scale on which parameterisation is necessary, there must be at least three tiers of variables. Then, the ‘true’ state of the system is defined by resolving all three tiers exactly. This can be quantitatively compared to both High Resolution models in reduced-precision that resolve the large- and medium-scale variables and parameterise the small-scale ones, and Low Resolution models in double-precision that resolve only the largest-scale variables and parameterise medium-scale ones.

This has motivated the introduction by the author (to be published in a paper currently in preparation) of a third tier of still faster-varying, small-scale variables labelled Z. The three tiers of variables are cyclically linked in space as shown in Figure 3, which shows the example where $K = 8$ X-variables are each linked to $J = 8$ Y-variables, which are each linked to $I = 8$ Z-variables. The relationship between the Y- and Z-variables is of a similar nature to that between the X- and Y-variables. The system was set up so that the scales X, Y and Z could be thought of as akin to large-

scale cyclones, convective clouds and small-scale turbulence respectively, although in the real atmosphere there is no such clean division between different spatial scales. Table 2 shows how the three-level system can be used to construct three models with different resolutions and levels of precision, which can each be compared against the 'truth' to test their accuracy.

Parameterising instead of resolving atmospheric phenomena at a given spatial scale yields considerable computational cost savings, so Low resolution models are less costly than High Resolution ones. However, because parameterisation schemes struggle to represent the true behaviour of the unresolved variables, Low Resolution models are also expected to be less accurate. This depends in part upon the nature of the parameterisation scheme. Following the method of Arnold et al (2013), in the present investigation of the three-level Lorenz '96 atmosphere so-called 'stochastic' parameterisation schemes are used. These incorporate random noise to represent the fact that the atmosphere will vary on scales smaller than those that the model can explicitly resolve, and that even very small perturbations will have an influence on its larger-scale evolution through the chaotic 'butterfly effect'. In each forecast, an ensemble of multiple parameterised runs is produced, each of which represents one possible effect of the unresolved phenomena. Stochastic schemes are often more accurate than non-stochastic alternatives (Arnold, et al., 2013) and play an increasingly important role in real-world models, which is what motivated their use in this investigation.

Using the models outlined in Table 2, it is possible to answer the key question that lies at the heart of this area of research: do High Resolution models in flexible precision predict the 'truth' more accurately than Low Resolution models in double-precision? If the answer to this question is 'yes', then assuming that both these types of model have similar computational costs, which would depend upon the inexact hardware used, flexible precision might offer a way of improving the accuracy of real-world forecasts.

4. Preliminary Results and Conclusions

This paper is intended to provide a first taste of what could become a revolutionary new approach to weather and climate forecasting, that of 'flexible precision' computing. If using flexible-precision inexact hardware in the supercomputers used to produce forecasts is to be beneficial to society, it must be capable of delivering more accurate forecasts for similar computational costs to conventional hardware. The aim of the experiments using the three-level Lorenz '96 was to emulate this hardware within an idealised atmospheric model and to test whether this was likely to be the case.

The full results of the investigation will be published in a forthcoming paper, but the preliminary findings are promising. In accordance with previous studies that have demonstrated large computational cost savings in both Lorenz '96 and less idealised atmospheric models with little degradation in accuracy using reduced precision, the technique was found to yield benefits for both long- and short-term forecasts of the state of the system. These preliminary tests compared the performance of Low Resolution models of the Lorenz '96 atmosphere in double-precision with that of High Resolution models with the variables represented in different degrees of flexible precision.

Short-term 'weather' and long-term 'climate' forecasting ability were analysed by comparing the models' output to the 'true' state of the system on these respective timescales. Computationally expensive High Resolution models with all variables represented in double-precision produced better forecasts according to both these metrics than Low Resolution ones, as would be expected. But, importantly, this advantage was not lost when using a much cheaper flexible precision High Resolution model that might be of similar computational cost to the Low Resolution one given appropriate hardware: changing the large-scale variables to single-precision and the medium-scale variables to half-precision had a negligible effect given the experimental error.

These exciting results emphasise the need for a multi-scale approach when attempting to make atmospheric models more efficient by reducing precision. If the results hold in real global models, flexible precision hardware with lower precision on smaller spatial scales could be used to improve forecast performance by increasing the model resolution without requiring much more computer speed and memory than is currently available. Further work, applying similar techniques to real-world global models, is planned to assess the scope of the gains that might be realisable in this way. It is also important for hardware manufacturers to be involved in creating the flexible-precision hardware needed to realise a revolution in the way weather and climate forecasts are produced.

Acknowledgements

The author would like to thank Tim Palmer and Peter Düben for initiating and inspiring the work upon which this paper is based and for providing invaluable help and advice. Thanks are also owed to Hannah Christensen for her insights regarding skill scores and stochastic parameterisation. The work was funded by the Natural Environment Research Council (NERC) and this paper is based on a presentation given by the author at the RMetS Student Conference in July 2015.

References

Arnold, H., Moroz, I. & Palmer, T., 2013. Stochastic Parameterisations and Model Uncertainty in the Lorenz '96 System. *Philosophical Transactions of the Royal Society A*, 371(20110479).

Duben, P. D. et al., 2014a. On the Use of Inexact, Pruned Hardware in Atmospheric Modelling. *Philosophical Transactions of the Royal Society A*, 372(20130276).

Duben, P. D., McNamara, H. & Palmer, T. N., 2014. The Use of Imprecise Processing to Improve Accuracy in Weather and Climate Prediction. *Journal of Computational Physics*, Volume 271, pp. 2-18.

Duben, P. D. & Palmer, T. N., 2014. Benchmark Tests for Numerical Weather Forecasts on Inexact Hardware. *American Meteorological Society Monthly Weather Review*, Volume 142, pp. 3809-3829.

Lorenz, E., 2006. Predictability - a Problem Partly Solved. In: *Predictability of Weather and Climate.* eds. T. N. Palmer & R. Hagedorn: Cambridge University Press, pp. 40-58.

Markov, I. L., 2014. Limits on Fundamental Limits to Computation. *Nature*, Volume 512, pp. 147-154.

Moammer, K., 2015. *Nvidia Pascal Launching in 2016 with 10X of Maxwell's Performance - Features 16nm, 3D Memory, NV-Link and Mixed Precision.* [Online] Available at: http://wccftech.com/nvidia-pascal-gpu-gtc-2015/ [Accessed 29 September 2015].

Palem, K. & Lingamneni, A., 2013. Ten Years of Building Broken Chips: The Physics and Engineering of Inexact Computing. *ACM Transations on Embedded Computing Systems*, 12(2).

Palmer, T., 2012. Towards the Probabilistic Earth System Simulator: A Vision for the Future of Climate and Weather Prediction. *Quarterly Journal of the Royal Meteorological Society*, Volume 138, pp. 841-861.

Palmer, T. N., 2014a. Build High-Resolution Global Climate Models. *Nature*, Volume 515, pp. 338-339.

Palmer, T. N., 2014b. More Reliable Forecasts with Less Precise Computations: A fast-Track Route to Cloud-Resolved Weather and Climate Simulators?. *Physical Transactions of the Royal Society A,* 372(20130391).

Wehner, M., Oliker, L. & Shalf, J., 2008. Towards Ultra-High Resolution Models of Climate and Weather. *The International Journal of High Performance Computing Applications,* 22(2), pp. 149-165.