REQUEST FOR A SPECIAL PROJECT 2020–2022

MEMBER STATE: GB

Principal Investigator[1]: Tim Palmer

Affiliation: University of Oxford

Address: Department of Physics
Atmospheric, Oceanic and Planetary Physics
Clarendon Lab. Parks Road
Oxford OX1 3PU

Other researchers: Matthew Chantry, Andrew McRae, Leo Saffin, Jan Ackmann, Milan Klower

Project Title: Imprecise approaches to accelerate weather forecasts

If this is a continuation of an existing project, please state the computer project account assigned previously.	**SP** GBTPIA	
Starting year: (A project can have a duration of up to 3 years, agreed at the beginning of the project.)	2020	
Would you accept support for 1 year only, if necessary?	YES ☒	NO ☐

Computer resources required for 2020-2022: (To make changes to an existing project please submit an amended version of the original form.)		**2020**	**2021**	**2022**
High Performance Computing Facility	(SBU)	15,000,000	15,000,000	15,000,000
Accumulated data storage (total archive volume)[2]	(GB)	15,000	15,000	15,000

Continue overleaf

[1] The Principal Investigator will act as contact person for this Special Project and, in particular, will be asked to register the project, provide annual progress reports of the project's activities, etc.

[2] These figures refer to data archived in ECFS and MARS. If e.g. you archive x GB in year one and y GB in year two and don't delete anything you need to request x + y GB for the second project year etc.

Principal Investigator: Tim Palmer

Project Title: Imprecise approaches to accelerate weather forecasts

Extended abstract

The completed form should be submitted/uploaded at https://www.ecmwf.int/en/research/special-projects/special-project-application/special-project-request-submission.

All Special Project requests should provide an abstract/project description including a scientific plan, a justification of the computer resources requested and the technical characteristics of the code to be used.

Following submission by the relevant Member State the Special Project requests will be published on the ECMWF website and evaluated by ECMWF as well as the Scientific Advisory Committee. The evaluation of the requests is based on the following criteria: Relevance to ECMWF's objectives, scientific and technical quality, disciplinary relevance, and justification of the resources requested. Previous Special Project reports and the use of ECMWF software and data infrastructure will also be considered in the evaluation process.

Requests asking for 1,000,000 SBUs or more should be more detailed (3-5 pages). Large requests asking for 10,000,000 SBUs or more might receive a detailed review by members of the Scientific Advisory Committee.

Motivation of the proposed research

ECMWF have set a bold and ambitious target of using 5km horizontal resolution for their ensemble forecasts by 2025, a significant increase on the current value of 18km. Meanwhile developments in high-performance computers have changed, with clock speeds stagnating but number of processors exploding. The inherently global calculations involved in integrating the Navier–Stokes equations pose a challenge to working efficiently in this massively-parallel paradigm. Reaching this target will therefore require improvements across the system. These challenges are by no means unique to ECMWF, weather and climate forecasters across the global are involved in major projects to work efficiently on current and future hardware.

We view our proposal as one cog in that machinery, reducing the cost of the forecast through two approaches. Weather and climate forecasts are inherently uncertain, with unresolved lengthscales and processes which need to be parameterised. These parameterisation schemes are missing information from unresolved lengthscales and therefore involve significant approximations. Ensemble forecasts also require stochastic perturbations to produce reliable forecasts where the spread matches the average error. With these factors in mind, calculating the whole system with high levels of precision should be considered a waste of processing power. It is this aspect that we believe can contribute towards the operational forecast target at ECMWF and guide work at other operational centres.

Our first approach will seek to remove unnecessary bits and carry out the majority of floating-point calculations at reduced numerical precision. This will continue our previous work which has already showed such reductions are possible across many components of the forecasting system. We seek to continue and expand upon this work. Below we outline our successes, particularly those working with OpenIFS, and discuss how we will progress this strategy.

The second approach will leverage developments in machine learning to build emulators for parameterised physics schemes. Using deep learning we will train emulators using data generated from the existing parameterisation schemes. These emulators will act as approximations to the current schemes while having lower cost. Work with other GCMs, mostly in simplified scenarios to-date, has demonstrated that these schemes can be learnt (O'Gorman and Dwyer 2018). Beyond accelerating the forecast, these neural network emulators could aid the data-assimilation process. Under the incremental 4D-var framework a tangent-linear and adjoint version of the code must be

maintained to assimilate observations. A major advantage of neural networks is their simplistic formulation, with the complexity lying in the weights of the networks. Deriving the tangent-linear and adjoint versions of a sufficiently accurate emulated parameterisation scheme would be an almost trivial task.

Summary of the most relevant work in previous studies in our group

Our work to-date has been to investigate reduced numerical precision across the kernels of the Integrated Forecast System (IFS). This work began with an assessment of the viability of single-precision for accurate weather forecasts in OpenIFS (Duben and Palmer 2014). This project was a major success, providing high quality forecasts at a 40% cost reduction and leading to the development of a single-precision mode of the IFS (Vana et al. 2017). Single-precision IFS is planned to be used for the operational forecast, meaning these cost reductions will be realised and the savings can be reinvested into resolution increases.

Following from the success of the single-precision IFS work, our group has focused on going below single-precision. There is increasing support for half-precision using hardware such as Nvidia's Volta GPUs or Google's Tensor Processing Units. We have investigated the feasibility to use lower-than-single precision for weather and climate forecasting. Our group created a software emulator to enable rapid testing of reduced precision kernels without the need to rewrite the code for the specialised hardware. Within OpenIFS we have examined three kernels to date. First, we investigated reduced precision in the spectral space calculations (Chantry et al 2019). Here, calculations are segregated by horizontal lengthscales, enabling the using of higher precision for the more predicable large-scales and lower precision for the smaller scales. We found that double or single-precision was necessary only for a small number of large-scale calculations while the vast majority could be calculated with half-precision or even lower.

Our second kernel was the Legendre transforms, one of the most expensive kernels in the IFS code. This kernel is dominated by matrix-matrix multiplications, which are a key component of neural networks. The Volta GPUs of Nvidia have been created for this purpose and can calculate half-precision matrix-matrix multiplications up to twelve times faster than a double-precision equivalent. Building on the scale-separation project described above we separate the transformations by lengthscale. Only the very largest scales are calculated with double-precision while the rest of the calculations are calculated with emulated half-precision. Taking this approach up to T1279, the operational forecast resolution, we are able to produce forecasts with comparable skill to a double-precision forecast when compared to the analysis (Hatfield et al 2019).

The third kernel investigated is the physical parameterisation package. As discussed in the motivation section these schemes are a large source of uncertainty and a major factor in both simulation time and number of lines of code. We have completed preliminary investigations using the reduced complexity GCM SPEEDY and T21 resolution forecasts with OpenIFS. In both of these settings we are able to carry out the vast majority of calculations at half-precision without significantly impacting forecast skill. We have also started a collaboration with Richard Gilham at UK Met Office to pool ideas and approaches to using reduced precision within parameterised physics across models.

Outside of OpenIFS, we have assessed reduced precision in data assimilation. Examining SPEEDY with an Ensemble Kalman Filter approach to data assimilation, it was found that half-precision was essentially sufficient. The remaining computational resources could be reinvested into additional ensemble members for an improvement in accuracy. In MITgcm, using gradient-based optimization, the adjoint model could be run in nearly half-precision without affecting the optimization process.

Beyond explicit reduced precision, we have also begun a body of work creating emulators of physical parameterisation schemes. To date we have created an emulator of the non-orographic gravity wave scheme from OpenIFS. To test these emulators in forecast mode requires interfacing machine learning tools (mostly written in python) and OpenIFS (written in Fortran). We have created routines in OpenIFS to import and run neural networks. We have carried out preliminary assessment of our best emulators when coupled into a T159 forecast. In these tests our forecasts with emulators produce comparable forecasts to those using the existing scheme. Currently our emulators have similar cost to the existing scheme are therefore work is required to improve their efficiency.

Projects that will be investigated in the next three years

Continue the testing of reduced precision physical parameterisation schemes

Tests with a simplified low-resolution model (SPEEDY) have been used to define steps for introducing reduced-precision parametrisations into a fully-complex model (e.g. OpenIFS). Initial tests with the single-column model of OpenIFS, using local resources, can be used to identify problems and biases that emerge at low precision and introduce changes to the code to optimise for low precision. Using the optimised code in the full OpenIFS, deterministic forecasts can then be run to identify the limit of precision before large errors start to emerge. With the special project resources, these deterministic forecasts can be run for various cases using a range of resolutions. With the optimal precision identified from the deterministic forecasts, this setup can then be verified by running ensemble forecasts using the special-project resources to allow us to have an acceptable ensemble size and a sufficient number of cases. This project will be led by Leo Saffin, a postdoctoral researcher in our group.

Create machine learnt emulators of physical parameterisation schemes

Preliminary work has been carried out to generate data-sets of two parameterisation schemes, orographic and non-orographic gravity wave drag. This data is currently being learnt using an array of deep neural networks to identify the optimal emulator, incorporating both accuracy and cost into this assessment of optimal. This has already been carried out for the non-orographic gravity wave drag scheme. Local resources will be used to test these emulators when coupled into OpenIFS at low resolution. Once satisfactory performance has been achieved for low resolution deterministic forecasts we will examine the performance for ensemble predictions and higher resolution forecasts. This analysis can use the tools for analysis of reduced-precision forecasts developed within our group.

In the later years of the special project will we use the knowledge gained examining the gravity wave drag schemes to create emulators of other parameterisation schemes. Our current objective schemes include the cloud and turbulent drag schemes. These schemes have higher complexity and cost, which could lead to increased savings if the project is successful. This project will be led by Matthew Chantry, a postdoctoral researcher in our group.

Examine the impact of reduced numerical precision in data assimilation

We also intend to continue our work using reduced precision for data assimilation. Recently, we have used the quasi-geostrophic (QG) model within ECMWF's Object-Oriented Prediction System framework and found that the tangent-linear and adjoint models can easily be run at single precision. Future directions could include QG experiments at much higher resolution, and the exploration of a large parameter space, such as different topographies or initial conditions. In meetings with ECMWF, it was decided that it would be impractical to test IFS's tangent-linear and adjoint models at single precision all in one step. However, a possibility is to investigate the effect

of single precision on individual linearised routines. This work will be led by Andrew McRae, a postdoctoral researcher in our group.

Reduced precision in elliptic solvers

Our group has also begun work examining reduced precision for elliptic solvers. These solvers are a critical kernel of grid-point models. For example, they are used at the UK Met Office, where single-precision has been tested in the preconditioner for the elliptic solver. At ECMWF, the IFS-FVM (finite volume model) is being developed as a rival to the spectral dynamical core by Piotr Smolarkiewicz. We have assessed the performance of the elliptic solver used by IFS-FVM in a low-resolution shallow water model and found excellent results even with the majority of the calculations performed at half-precision. Future work in this direction would use the special project units to examine the effect of reduced precision at higher resolutions and for models with increased complexity. This project will be led by Jan Ackmann, a postdoctoral researcher in our group.

Physical parametrisation emulators in data assimilation

In our motivation we outlined another possible benefit for machine learnt emulators of physical parameterisation schemes, the ease of generating tangent-linear and adjoint versions for use within 4D-var. Operational 4D-var at the UK Met Office uses reduced complexity physics in data assimilation to limit the upkeep and running costs of this component. Providing our work above, generating emulators of the parameterisation schemes, is successful, we will generate the tangent-linear and adjoint versions of the neural networks and deploy them in the data assimilation mode of IFS. This project will be led by Matthew Chantry, a postdoctoral researcher in our group.

Justification of computer resources and technical characteristics

We plan to carry out simulations of OpenIFS with replacements to various kernels. These replacements will be either using the existing algorithm calculate at reduced precision or use a machine learnt replacement of the existing scheme. Our current work has used cycles 38 and 40 but we plan to migrate to the new version of OpenIFS, cycle 43r3, once released. This new OpenIFS cycle will introduce the cubic-octahedral grid and single-precision options. For both approaches similar testing scenarios will be necessary. Initially both approaches will be tested using individual simulations at TL159 (or TCo199) resolution for a small number of dates. These simulations will establish that the chosen precision or machine learning replacement produce plausible forecasts. The cost of these initial runs will be small compared the main assessment. The full testing will involve large ensembles across multiple start dates. For twelve start dates and 25 ensemble members of TCo199 this would cost approximately 90,000 System Billing Units (SBUs). For the same assessment of TCo399, this would cost approximately 420,000 SBUs. For reduced precision experiments the reduced precision emulator significantly decreases the performance of those parts of the code being calculated at reduced precision. This performance decrease can increase the cost of the experiments by up to a factor of 20, depending on the original computational cost of the kernel containing our emulator. For spectral space calculations studied in the last special project the effect was a factor 10 increase in cost. Similar factors here applied to the above experiments would then cost 900,000 or 4,200,000 SBUs respectively for TCo199 and TCo399. In the case of the neural network emulators the costs would not be expected to exceed the normal SBU.

For our proposed budget 15,000,000 SBUs we could carry out ten TCo199 ensemble assessments with reduced precision kernels and ten TCo399 ensemble assessments with neural network emulators each year. Depending upon the successes of the various projects the balance between these two elements might shift. Our existing knowledge of OpenIFS and the ECMWF supercomputing facilities means we can commence with significant simulations in the first year and do not need an increased number of SBUs in the second and third years of the project.

References

O'Gorman PA, Dwyer JG. Using Machine Learning to Parameterize Moist Convection: Potential for Modeling of Climate, Climate Change, and Extreme Events. Journal of Advances in Modeling Earth Systems. 2018 Oct;10(10):2548-63.

Düben PD, Palmer TN. Benchmark tests for numerical weather forecasts on inexact hardware. Monthly Weather Review. 2014 Oct;142(10):3809-29.

Váňa, F., Düben, P., Lang, S., Palmer, T., Leutbecher, M., Salmond, D. and Carver, G., 2017. Single precision in weather forecasting models: An evaluation with the IFS. *Monthly Weather Review, 145*(2), pp.495-502.

Hatfield S, Chantry M, Düben P, Palmer T. Accelerating high-resolution weather models with deep-learning hardware. In Proceedings of the Platform for Advanced Scientific Computing Conference 2019 Jun 12 (p. 1). ACM.