Implementing the Operational Weather Forecast Model of MeteoSwiss on a Cray XT3

Tricia Balle, *CSCS and* **Neil Stringfellow**, *CSCS*

ABSTRACT: *The Swiss weather service, MeteoSwiss, uses the computer facilities of the Swiss National Supercomputing Centre (CSCS) to carry out the twice daily weather forecasts for Switzerland using a specialized alpine model called aLMo, based on the LM code from the COSMO consortium. This paper describes the implementation of the aLMo suite on the Cray XT3 at CSCS and outlines some of the issues faced in porting and configuring a complex operational suite on a new architecture. Performance figures for the suite are given, and an outline is presented of some of the challenges to be faced in carrying out the next generation of high resolution forecasting.*

KEYWORDS: Cray XT3, XT4, operational weather forecasting, COSMO

1. Introduction: MeteoSwiss and the Swiss national forecasts

"MANNO, Switzerland and SEATTLE, WA, March 12, 2007 -- In February 2007, MeteoSwiss began production weather forecasting using a Cray (Nasdaq GM: CRAY) supercomputer located at the Swiss National Supercomputing Centre (CSCS). This is a key transition to a scalable supercomputing technology that will allow MeteoSwiss to continue improving forecast quality through model and resolution improvements. Plans to implement higher-resolution forecasts in January 2008 will make detailed forecasts of Switzerland's intricate Alpine topography possible for the first time. This will make Switzerland one of the first countries in Europe to move from the current standard forecast resolution of seven kilometers to a two kilometers resolution."

On December 26 1999 a low-pressure system called Lothar resulted in a violent extratropical cyclone that swept rapidly across central Europe and caused major damage in France, southern Germany, and Switzerland, with wind speeds of around 150km/h in lower areas and over 250 km/h on some mountains. Not only buildings and infrastructure, but also forests such as the Black Forest sustained substantial economic losses (US$12.8 billion; 3% of Swiss timber reserves were damaged). Some weather services were criticized for not issuing a storm warning for Lothar. The storm began as a frontal wave over the Western Atlantic and developed rapidly in less than 24 hours, and it moved across Switzerland in two and a half hours. Although the storm was extremely strong, it was within the bounds of what is expected on average every 10 to 15 years and so better forecasting models able to capture such events are essential.

MeteoSwiss is the Swiss national weather service, and one of its primary functions is to provide accurate forecasts for periods of up to a few days ahead using techniques that rely heavily on computational methods of numerical weather prediction. These predictions are carried out using the facilities available at the Swiss National Supercomputing Centre (CSCS); up until February 2007, the main computational engine for the twice-daily operational forecasts was an NEC SX-5. However, in February 2007 the operational forecasts were moved to a Cray XT3; within the next year, the forecast resolution will be increased and the forecasts will be run more frequently on a Cray XT4. CSCS is the first site in the world to run operational weather forecasting on a Cray XT3.

The software used to carry out the forecast simulations is version 3.16.2 of the LM code developed by the German weather service, Deutscher Wetterdienst (DWD) as part of the Consortium for Small-scale Modeling (COSMO), and is a local model, which allows a relatively simple subdivision of the particular part of the Earth that is of interest without requiring special programming to consider effects at the planet's poles.

For the aforementioned future operational model, the contract between CSCS and MeteoSwiss specifies the time to solution of a finer resolution operational forecasting suite to be run eight times a day, as well as requiring a number of on-demand emergency simulations. Reliability criteria are laid down for the scheduled operational simulation cycles based on the number of delayed or incomplete simulations per calendar month, and 95% of the on-demand tasks have to be successfully completed "on time" per calendar year. There is also a requirement for storage (one simulation cycle produces about 42GB of data): data archived over the last 365 days must be accessible in not longer than a couple of minutes, and security against accidental loss of data must be provided.

Current Operational Model

The current operational model used by MeteoSwiss, now running on a 2.6Ghz dual core Cray XT3 instead of the NEC SX-5, consists of two aLMo/7 (lower resolution) runs per day, each a 72-hour simulation, with a grid size of 385 x 325 points with 45 atmospheric levels and a time step of 40 seconds. Comparing the performance of the SX-5 and the XT3, the elapsed time for a 72-hour simulation is:

NEC SX-5	14 cpus (dedicated)	78 mins
Cray XT3	50 dual core nodes	73 mins

With a number of additional optimizations specific to the XT3 (described later), the elapsed time for the XT3 drops to 66 minutes.

Although the region used by this model is very similar to the area covered by the operational model of DWD, this resolution employed by MeteoSwiss is lower but there are more atmospheric levels used in order to provide better coverage of local effects in the Alpine region. The specific alpine model is able to take account of some of these effects, but better modeling of some aspects of the weather in mountain regions cannot be covered at that resolution.

Future Operational Model

The future operational model will no longer simply carry out two simulations per day, each consisting of a 72-hour forecast over the whole of Western Europe. Instead, MeteoSwiss will move to a configuration where these forecasts are supplemented by aLMo/2 (higher resolution) 18-hour forecasts over the Alpine arc at 3-hour intervals. The most expensive simulations are required to complete two assimilation runs together with one 18-hour aLMo/7 low resolution forecast and one 18-hour aLMo/2 high resolution forecast; combined with the necessary interpolations and time-critical postprocessing, this must take less than 30 minutes real time according to the contract. These higher resolution runs will start in the middle of this year and will be operational by January 2008.

Note that the 72-hour aLMo/7 forecast will still be run twice daily in addition to the 8-times daily COSMO 2K suite, as the future operational suite will be called from now on.

In the current model the assimilation cycles, where data is taken from external sources and prepared for input to the LM code, are carried out separately from the numerical simulations at different times of the day, but in the upcoming model these are incorporated into a chain of seven programs which work together, and therefore the time criticality of the results from these runs has increased. The move to this higher resolution model with more frequent forecasts over a smaller domain to supplement the regional model allows several advantages over the previous operational suite and should provide benefits for Switzerland in terms of the ability to provide warnings based on the prediction of local events such as floods or avalanche danger, the ability to enhance national security with respect to nuclear power plants, and the ability of MeteoSwiss to engage in closer collaboration with international partners engaged in high resolution short-range modeling.

A benefit of the new model is that it should allow for the capture of extra features that are not able to be modeled at the regional level, for instance, by enabling better account to be taken of local topography; in particular, it should allow for the resolving of Alpine valleys to a high resolution, which is generally a huge problem when attempting to forecast in this geographical region. As Marie-Christine Sawley, co-Director of CSCS, mentioned in a recent press release about MeteoSwiss forecasts run on the XT3, "…the mountainous terrain can result in discontinuous weather patterns. We can have a destructive storm in one valley while the sun shines in the neighboring valleys." Further code developments should also allow for the modeling of extra physical effects that are not currently present, such as summer convection. The use of frequent short-range forecasts should also enable the possibility of some degree of ensemble forecasting.

One important aspect of the move to the new operational suite is that whereas an increase in computational power within a computer center is normally used to carry out simulations on larger problem sizes, COSMO 2K attempts to solve one problem of the same size together with a slightly larger problem in a much shorter time than the original; as is well known, the normal rules of scalability on parallel systems show that this is a much more difficult task to accomplish.

An additional demand placed on the facility is that the chaining of the parts of the operational cycle may require extra time between the assimilation runs and the aLMo simulations, as job completion may not be immediately acted upon by the operating system to trigger the dispatch of the next executable; this can be particularly true on multinode systems where there is also a higher possibility that one part of this chain leads to a node suffering a software problem, which could cause the whole operational cycle to be delayed.

A further change in moving to COSMO 2K is the dramatic effect it will have on the remaining usage of the system, with interruptions planned every 3 hours for a 30-minute interval. This interruption level means that other scientists who wish to run jobs on the same system have to submit simulations requiring less than two and a half hours of compute time, a situation which is likely to be unacceptable to many researchers and certainly leaves large gaps in the schedule. Furthermore, this does not take into account any effect on other researchers of the emergency on-demand usage described below.

For the reasons mentioned above, CSCS is planning to migrate the MeteoSwiss operational forecasts onto a dedicated machine so that the forecasts do not interfere with other users; a 5 cabinet Cray XT4 will installed at CSCS within the next few months in order to run the new forecasts. This will be discussed in more detail in section 3.

Note on unscheduled on-demand usage. In addition to the regular COSMO 2K suite, the contract with MeteoSwiss also provides for 1) the running of a dispersion suite consisting of the LPDM (Lagrangian particle dispersion model) code, typically once or twice a week, each run equivalent to about a 2 hour aLMo/2 forecast; and 2) the running of the COSMO 2K suite in order to calculate a rapid update cycle with 1 hour refresh, at most for 24 hours a year in two episodes. The contractual obligation is for a maximum of 5 minutes to elapse between the triggering of the latter demand on the XT3 and its effective start; such a mandated fast response would clearly impact adversely on any other users of the system.

2. The COSMO 2K suite and the operational model

The LM Model

The LM-RAPS code is the RAPS version of the LokalModell (LM) of DWD (COSMO), which is a numerical weather prediction code used for regular forecasts by several national centers throughout Europe including the twice-daily forecasts issued by MeteoSwiss.

When using global forecasting codes, special consideration needs to be given to modeling at the north and south poles, or at other singular points where a spherically mapped coordinate system generates problems for a rectangular grid system. With a local model, no special effects need to be considered at the poles, and furthermore codes such as the LM allow a transformed coordinate system by defining a new pole so that the region of interest is not distorted, and in this case it is possible to use the code to carry out a forecast at the true north pole by defining a new pole for the model which lies on the equator, thereby ensuring that the grid over the true north pole is as close to rectangular as possible.

For both the low and high resolution simulations in the operational suite, the pole is placed at a point with a latitude of 32.5° North and a longitude of -170° so that the virtual equator passes through Bern and the region of interest shows little distortion due to the curvature of the Earth.

To help in the analysis of the results of the simulation and to ensure the correct execution of the code it is possible to have the code produce a set of output files that begin with the letters ”YU” and which give details such as timing information, grid distribution between processors, and a sample of some important values to allow verification of the numerical results of the simulation. Some of these files can be used for verification of results and for determining timing information of various parts of the code.

The execution of the code is driven by a set of Fortran namelists that are given in files with predefined names, and these namelists are used to provide details such as the location of input data and where to place output files, as well as information on which alternative parts of the code to execute such as the choice of different communication schemes as well as which numerical schemes to use, for example, whether to select a simple time-stepping leapfrog scheme or to use a Runge--Kutta method. One important value given in the RUNCTL namelist is the value of nboundlines, which determines the size of the halo region used in the simulation and, more importantly, when considering the distribution of grid points, the specification of this value also reduces the region on which calculations are to take place by 2*nboundlines columns and rows.

The Alpine Model

The current alpine model employed by MeteoSwiss takes input boundary conditions from a global model calculation carried out by the European Centre for Medium Range Weather Forecasting (ECMWF). As mentioned above, the forecast is for a 72-hour period, and this will still be used for the daily weather forecasts but supplemented by higher resolution 18-hour forecasts for emergency planning and modeling of local phenomena that cannot be captured at the lower resolution.

Current model (aLMo/7)

The grid for the aLMo/7 model has a size of 385 x 325 points covering an area of Western Europe centered over Switzerland, with a resolution of one sixteenth of a degree (which translates to approximately 7 km between adjacent grid points in the centre of the region). In the vertical dimension, the model has 45 atmospheric levels covering a vertical range from the ground to a distance of approximately 25 km, with a higher concentration of these levels in the lower part of the atmosphere. The distribution of points to be calculated utilizes a smaller range of 381 x 321 points with the remaining outer points acting as ghost cells or halo cells, which are primarily used to provide boundary conditions for the calculations at the other grid points. For this resolution a time step of 40 seconds is sufficient to capture the necessary physical phenomena with the model employing a straightforward time stepping scheme.

Higher resolution model (aLMo/2)

For aLMo/2 the grid size will be increased to use 520 x 350 points with 60 atmospheric levels (the number of levels was determined by carrying out various simulations in the early stages of testing of aLMo/2). The resolution of aLMo/2 is one fortieth of a degree (which translates to approximately 2.2 km in the centre of the region and is therefore a considerably higher resolution than the aLMo/7 regional model). As described before for the aLMo/7 configuration, the actual number of points to be calculated works on a smaller grid of 516 by 346 points with halo cells. At this resolution the time step has to be reduced to 18 seconds, and use is now made of a Runge--Kutta scheme for numerical integration.

This higher level resolution has required a great deal of work in refining the model to take account of new physical phenomena that only exhibit themselves at this more local scale, and for this reason the full aLMo/2 suite is not due to be used in an operational mode until the beginning of 2008. An example of the new challenges faced by changing the resolution of these models is an instability which occurred in Greenland valleys in some simulations carried out by researchers at the UK MetOffice, where trapped cold air led to the model crashing; more relevant to our situation, this instability also exhibited itself when carrying out a 2 km resolution model as part of a MAP case study in the Val d'Aosta region of Italy, which forms part of the Alpine arc.

The COSMO 2K suite

The time-to-solution criterion (30 minutes including time-critical postprocessing, or 25 minutes for the HPC segment of the suite) is based on a benchmark suite run and validated at CSCS consisting of interpolation, assimilation, and forecast runs at various resolutions. The main characteristics of the benchmark were as follows.

(i) aLMo/7: mesh of 385 x 325 points (or 285 x 225 points), with 60 levels and a time step of 72 simulated seconds. Boundary conditions are applied at 3 simulation hour intervals, with Rayleigh damping at the top. The numerical solver used is Runge--Kutta, all possible dynamical fields are calculated, and the radiation routine is called every 60 simulated minutes. The GRIB output files are produced every 60 simulated minutes.

(ii) aLMo/2: mesh of 520 x 350 points, with 60 levels and a time step of 18 simulated seconds (15 seconds for the assimilation runs). Boundary conditions are applied hourly, and again there is Rayleigh damping at the top. As for aLMo/7, a Runge--Kutta solver is used and all possible dynamical fields are calculated, with the radiation routine called every 30 simulated minutes. There is no parameterized deep convection, and the output GRIB files are produced every 30 simulated minutes.

The full benchmark then consisted of the following seven parts, run in sequence.

(i) Interpolation from IFS to aLMo/7: 385x325 x 60 levels
(ii) aLMo/7 assimilation: 385x325 x 60 levels, 72s timestep, 6 hour simulation (depends on (i))
(iii) Interpolation from IFS to aLMo7: 285x225 x 60 levels
(iv) aLMo/7 short forecast: 285x225 + 60 levels, 72s timestep, 18 hour simulation (depends on (iii))
(v) Interpolation from aLMo/7 to aLMo/2: 520x350 x 60 levels (depends on (iv))
(vi) aLMo/2 assimilation: 520x350 x 60 levels, 15s timestep, 3 hour simulation (depends on (v))
(vii) aLMo/2 forecast: 520x350 x 60 levels, 18s timestep, 18 hour simulation (depends on (v))

Note that in practice it is not necessary to run all parts in sequence – for example, once interpolation (iii) has commenced, then the corresponding aLMo run (iv) can be started before the interpolation has finished.

3. Porting of the LM code to the Cray XT3

The aLMo/2 model size was determined after considerable preliminary work performed at CSCS over the past two years to analyze the code on a number of proposed platforms. A detailed benchmarking exercise was performed by CSCS in 2005 on an SGI Altix 3700, the NEC SX-8, the Cray XT3 and X1, and an AMD Opteron cluster with Quadrics interconnect.

The Cray XT3 at CSCS was selected in a joint project with the Swiss research centre the Paul-Scherrer-Institute (PSI). As of May 2007, there are two Cray XT3 machines at CSCS; the largest now consists of 1664 2.6Ghz dual core compute processors and a total of 3.3 Terabytes of memory. There is also a smaller 2.6Ghz dual core machine with 74 dual core nodes, which is where the MeteoSwiss production runs are now located.

The LM code is currently compiled on the Cray XT3 using version 6.2.5 of the PGI compiler. During the initial port, a compiler bug was identified so that one (unimportant) routine has to be compiled with no optimization, but this is fixed in a later compiler. Calls to system_clock were replaced by calls to mpi_wtime.

The major problem in porting the software to the Cray XT3 was in finding an appropriate method of implementing the GRIB library, which is used for file I/O. The GRIB library uses a number of calls not available on Catamount in order to check file access parameters and to place locks on files, and it was necessary to introduce conditional compilation directives to remove these elements of the code from the version for the XT3.

Work was done both by Cray and CSCS to bring the operational COSMO 2K benchmark suite described above to within the time-to-solution criterion. The following timings are from a single core 2.6Ghz XT3. The main aLMo runs were made on 654 single core processors (26x25 decomposition plus 4 I/O nodes); the shorter interpolation runs were made on fewer processors as they did not scale much beyond 100 processors.

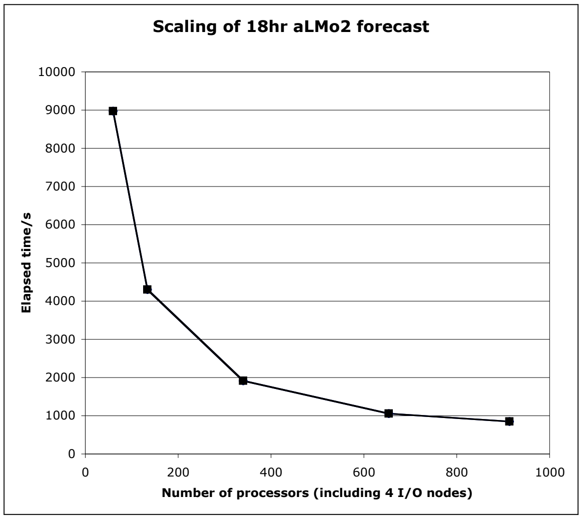
	Walltime	Decomp
(i) ifs2lm	0:11.70	(8x8+1)
(ii) aLMo/7 assimilation	1:57.44	(26x25+4)
(iii) ifs2lm	0:14.42	(8x8+1)
(iv) aLMo/7 short forecast	2:18.91	(26x25+1)
(v) lm2lm	1:25.96	(10x10+1)
(vi) aLMo/2 assimilation	3:26.25	(26x25+4)
(vii) aLMo/2 forecast	17:09.75	(26x25+4)

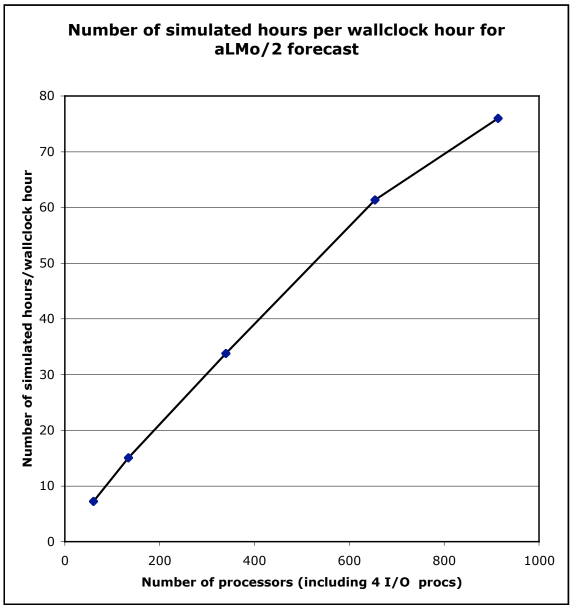
Total time to solution: 26.7 minutes

Scaling of fine resolution forecast run (18 hour simulation)

The following runs were made on a heavily loaded single core Cray XT3 and are not the fastest possible for each node count, but this scaling is representative. The two graphs show the elapsed wallclock time for the various processor counts, and also the scaling of the number of simulated model hours per wallclock hour.

Number of cpus	Decomposition	Elapsed time/s
60	8x7+4	8973.95
134	13x10+4	4304.83
340	24x14+4	1918.39
654	26x25+4	1076.19
914	26x35+4	852.85





Code modifications for optimization

Very few significant changes had to be made to the source in order to meet the required time to solution. The PGI optimization flag –fastsse was used throughout the compilation. Some repeated closing and flushing of selected YU* text output files was removed, which avoided some significant performance slowdowns yet did not affect the code functionality. A namelist option is to be added to the main LM source branch in order to enable this functionality in the future. A number of (unnecessary) target attributes were removed from variable and array declarations. The reason for this is that the PGI compiler does not fully optimize loops containing arrays with the target attribute. This is understandable, but unfortunately, unlike some other compilers, there is no way to inform the PGI compiler that it is safe to optimize the loop (for example, via a compiler flag guaranteeing no hidden equivalences and stride-1 accesses). Removing the target attributes can only be a workaround until the compiler is modified. Finally, a call to gather_values (which does an mpi_gather) was replaced by a standard call to mpi_alltoallv in order to transfer the data to be written to file; the alltoall call means that only the relevant part of each field is sent to each processor and this communication can be performed in parallel. This last modification is most effective at high processor counts.

A change was made to copen.c, in the GRIB library, in order to allow IOBUF buffering to be used in conjunction with GRIB files.

Runtime modifications and detail

The LM code can be run with a varying number of dedicated asynchronous I/O processors. After some experimentation, it was decided to run most of the aLMo assimilation and forecasting runs using 4 dedicated I/O processors. It was also found that the default MPI rank ordering, which ensures that a maximum of one I/O processor resides on a single node, was optimal for these runs.

It was important to set ltime_barrier to .FALSE. in the input namelist in order to avoid unnecessary time spent in timing barriers.

The I/O buffering utility IOBUF was used to buffer some of the larger input and output GRIB files, though care has to be taken not to buffer too many files on smaller memory systems to avoid memory exhaustion.

For the timings above 1MB lustre stripes across 1 or 2 OSTs were used, though more investigation into striping is ongoing.

The -small_pages flag for yod was used for all runs.

Single to dual core

Moving from single to dual core, it was found that 862 2.6Ghz dual core XT3 processors gave the performance equivalent to 654 single core processors. In general we see a penalty of 20 to 25% when moving from single to dual core. A good part of this penalty is due to the logic in the fast waves Runge-Kutta section of the code; this was introduced along with the high-resolution logic of the LMK model, which supports the very low distances between grid points. The fast waves logic is heavily reliant on memory bandwidth and very cache unfriendly. In order to improve the cache utilization of this part of the code, a significant rewrite will be necessary and no good solution has yet been suggested. However the logic may now be mature enough to approach the problem.

Moving to the Cray XT4

CSCS will install a 5 cabinet Cray XT4 (2.6Ghz dual core processors, 2GB compute nodes, 120 total blades) to run the future MeteoSwiss operational suite. This machine should be installed by the summer so testing can begin as soon as possible. The benchmark suite has been run on an XT4 configured similarly to the projected CSCS machine and various grid-size configurations have been compared. The table below shows the best elapsed time observed for each configuration over a number of runs on an empty and loaded machine. Clearly, the XT4 will be able to meet the time-to-solution target.

Grid Size	Number of Cores	Number of Blades	Elapsed time (mins)
26x35+2	912	114	23:55
25x35+4	879	110	23:51
20x43+4	864	108	24:26
12x69+4	832	104	27:19

It can be seen that the preferred solution will be to run the suite on 110 blades in 23:51 minutes, although it is also possible to fall back to 108 blades if necessary in a situation where the machine suffers an abnormally large number of compute node failures. In this case, only 5 lines of input script need to be altered, so this should not cause any problems. Assuming no failures and assuming 5 service blades out of the 120 total, running 110 blades for the suite leaves 5 spare compute blades, which is a good buffer. In the worst-case scenario, when the suite is unable to run at all on the XT4, a backup solution will be to run on the existing 18-cabinet XT3. After the April 2007 dual core upgrade, this machine has 3328 compute cores, 1100 of which will be sufficient to run the COSMO 2K suite in under 25 minutes.

Another advantage of leaving 5 blades free is that two or three or them could be configured as service blades with 8GB/node, each running standard Linux. They would then be able to run the post-processing work that is currently offloaded to another platform; this has various advantages, including the fact that the output data from the HPC runs is immediately available to the post-processing suite without the necessity for data transfer. Running the full COSMO 2K suite on a single platform also reduces the number of points of failure when diagnosing problems, and means that all parts of the suite have full vendor support.

4. Conclusions

Over the past two years, CSCS has worked with MeteoSwiss and Cray Inc. to successfully migrate the Swiss national forecast from the NEC SX-5 to the Cray XT3, and is currently the only site worldwide to run operational weather forecasts on an XT3. Looking to the future, by the end of the first quarter of 2008 both the current lower resolution forecasts along with a newer higher resolution forecast suite will be run on a Cray XT4. The resulting ability to better forecast weather patterns over the Alpine region should mean that events such as Lothar are far less likely to occur without prior warning.

Acknowledgments

A number of people aside from the authors have been involved in the analysis, porting, and optimization of the COSMO 2K suite on the Cray XT3 and XT4. Among them are Roberto Ansaloni and Eckhard Tschirschnitz of Cray Inc., Emanuele Zala, Jean-Marie Bettems, and Guy de Morsier of MeteoSwiss, and Mark Cheeseman, Angelo Mangili, and Mauro Ballabio of CSCS.

About the Authors

Tricia Balle is an HPC benchmarking and applications contractor working most recently with both CSCS and Cray Inc. She can be reached at PO Box 28, Tuakau, New Zealand (email tricia@auspira.com). Neil Stringfellow is Senior Applications Analyst and ALPS Programme Manager at CSCS. He can be reached at CSCS, Galleria 2 – Via Cantonale, CH-6928 Manno, Switzerland (email nstring@cscs.ch).