Objective calibration of weather prediction models

Dr. Antigoni Voudouri, Hellenic National Meteorological Service (HNMS), Greece, PI
Dr. Jean-Marie Bettems, Federal Office of Meteorology and Climatology MeteoSwiss, Co-PI
Dr. Pavel Khain, Israel Meteorological Service (IMS), Israel, Co-PI

Project team members:
Dr. Marco Arpagaus, Federal Office of Meteorology and Climatology, MeteoSwiss
Dr. Omar Bellprat, Catalan Institute of Climate Sciences (IC3), Spain
Dr. Oliver Fuhrer, Federal Office of Meteorology and Climatology MeteoSwiss
Prof. Christoph Schär, Institute of Atmospheric and Climate Science, ETH Zurich

Abstract

All atmospheric models used for numerical weather prediction (NWP) and climate modeling have inherent uncertainties. Many of them stem from parameterization schemes for physical processes within the models, which often include free or poorly confined parameters. Model developers normally calibrate the values of these parameters manually in order to improve the agreement of forecasts with available observations. This 'expert tuning' is typically done once during the development of the model, for a certain target area, and for a certain model configuration, and is often difficult if not impossible to replicate. It is questionable whether such a calibration is still optimal for different target regions (e.g. with a different climate) or for other model configurations (e.g. with an increased grid resolution). Furthermore, the lack of an objective process to re-calibrate the model is often a major roadblock for the implementation of new model features.

A practicable objective multi-variate calibration method has been developed by Bellprat et al. (2012a and 2012b) and implemented for a regional climate model. The objective method has shown to be at least as good as an expert tuning. Based on these results, a research project (CALMO) has been proposed and accepted with the aim to investigate how to transfer this method to NWP applications. Within the framework of the CALMO project, **a total of 1'070'000 node hours are requested on the hybrid Piz Daint system at CSCS to develop an objective calibration method for NWP.** As a demonstration vehicle, we will use a new kilometric configuration of the COSMO model. Since many research groups and operational centers are moving towards (convection-permitting) kilometric resolutions, there is a particular interest for re-calibrating high-resolution configurations. At the same time, there is a high potential to show a significant impact of the calibration method, since the kilometric configuration differs substantially from the COSMO configurations widely used. The need to span a significant subset of the model parameter space and the size of the computational mesh requires access to significant computing resources. The Piz Daint system is optimal for this purpose, as we are planning to run the COSMO model in GPU mode.

Besides setting up the calibration method for an NWP model, the two additional scientific goals of this project are to understand the sensitivity of the NWP model quality with respect to the model parameters space as well as to optimize the calibration procedure in order to minimize the amount of computing resources required. Both aspects cannot directly be transferred from the experience with the calibration of the climate version of COSMO due to the very different performance scores and forecast lengths used for climate and weather forecasts.

The main scientific impact of a positive outcome of this project is the availability of an objective calibration tool to determine the optimal setting of free or poorly defined model parameters. Depending on the (minimal) computing resources needed for a robust calibration, modelers will be able to objectively and reproducibly re-calibrate their NWP modeling system whenever needed: after major model changes, for an unbiased assessment of different modules (e.g. parameterization schemes), to avoid or remove compensating errors, for optimal perturbation of parameters when run in ensemble mode, for a better understanding of the sensitivity of the model quality to a specific model parameter, etc. What today is only done once ('expert tuning') will in the future be done as often as needed!

Background and significance

In this section, we briefly outline the current state-of-the-art of objective calibration for atmospheric models and describe the methodology we will apply for our calibration.

It has been shown that model parameter uncertainty is a major source of errors in ***regional climate model*** simulations (Stephens et al., 1990; Knutti et al., 2002; Webb et al., 2013). To circumvent this problem, an objective calibration method (Neelin et al., 2010) for the climate version of the COSMO model, the so-called COSMO-CLM, has been applied at the Institute of Atmospheric and Climate Science of the ETHZ (Bellprat et al., 2012b). After having identified key COSMO ***model parameters*** (Bellprat et al., 2012a) and defined a ***performance score*** representative for the model quality, a cost-effective ***meta-model*** describing the model performance in the space spanned by these model parameters has been derived. The optimal parameter configurations for the full model are then found by optimizing the model performance of the meta-model with respect to the performance score used.

The calibration performed by Bellprat et al. (2012b) allowed for the reduction of the model error of an expert tuned COSMO-CLM by about 10% using at the same time much less human resources. The optimal parameter setting was also found to be close to the COSMO-EU configuration, the configuration used in production for NWP forecasts at the German Weather Service (DWD), suggesting a low dependency of the calibration results with respect to the specific application of the model (climate or weather). In the published version of the COSMO-CLM calibration (Bellprat 2012b), a total of 5 parameters were calibrated. In a second calibration attempt that is currently being prepared for publication, three additional parameters were considered, leading to a significant further improvement. In particular, the warm summer temperature bias (and the overestimation of inter-annual summer temperature variability) has been strongly reduced. It is worth noting that this summer temperature bias is a persistent bias of the COSMO-CLM, which has resisted all previous expert tuning attempts (see Kotlarski et al., 2014).

The basic idea of the calibration framework is to build a computationally efficient statistical model that approximates the model output fields of the full model for an *n*-dimensional model parameter space (often termed as model emulator, statistical surrogate model or *meta-model*). A model emulator (O'Hagan, 2006) allows estimating the simulated model variables of the climate or weather prediction model of interest for a certain set of model parameters selected.

There are numerous approaches how to approximate model parameter experiments (e.g. Neural Networks or Gaussian Processes) and several approaches have been applied to climate models as discussed in Annan and Hargreaves (2007). The proposed project relies on a meta-model that approximates the parameter space using a multi-variate quadratic regression. This choice has been made initially for tuning COSMO-CLM as the approach uses only the minimum number of model simulations that are required to account for non-linear behavior and parameter interactions in model parameter experiments. The use of a quadratic regression further inhibits over-fitting and allows for analytical solutions of the parameter space.

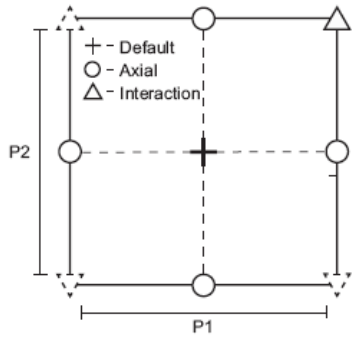

Figure 1: Koshal parameter experiment design required to estimate the parameters of the meta-model (from Bellprat et al., 2012b)

In order to estimate the coefficients of the multi-variate quadratic meta-model, a minimum set of model runs has to be performed. The most obvious way of carrying out the minimum set of runs is a Koshal design for fitting a second-order polynomial (Mayers et al., 2009). Such a design is illustrated in Figure 1 for two parameters (P1, P2), where the default value is in the center (cross) and the minimum and maximum values are sampled in each dimension (circles). Additionally, in order to take

parameter interactions into account, one corner point needs to be simulated (triangle). Using this design a total number of N*2+N(N-1)/2 simulations are required for a calibration of N parameters. So for a calibration of 8 parameters, a total number of 44 simulations are required.

State-of-the-art ***NWP models*** are tuned using expert knowledge and hand-tuning without following a well-defined strategy (Duan et al., 2006; Skamarock, 2004; Bayler et al., 2000). This 'expert tuning' is typically made only once during the development of the model, for a certain target area, and for a certain model configuration, and is difficult if not impossible to replicate. To use an objective method such as the one applied in Bellprat et al. (2012b) is highly attractive due to its efficiency, wide calibration range and transparency. A re-calibration of the model parameters could and indeed should be applied each time a significant change in the configuration is introduced, or when the model is used on a target region with a significantly different climatology. Model development could thereby be accelerated, because the expert knowledge required for an expert tuning is often not readily available, and testing new model parameterizations would ideally be accompanied a proper re-calibration. Additionally, a major stumbling block to model improvements are compensating errors, where the systematic error in a certain part of the model is compensated by manual tuning of another part of the model and thereby introducing another systematic (but balancing) error. As a result, compensating errors often lead to a degradation of model quality if a significant improvement is made to the model component with the systematic error. An automatic re-calibration methodology can help to surmount this deadlock by being able to rapidly find new optimal parameter settings.

The calibration method proposed by Bellprat et al. (2012b) for regional climate modeling cannot be directly applied to NWP. First and foremost, the performance score used to assess model quality is not applicable to NWP model calibration and a new performance score has to be developed. Secondly, the length of the NWP model integrations is much shorter (days) than regional climate model integration (years). Thus, there is considerable potential to optimize the simulation strategy with respect to the minimal amount and distribution of NWP model forecasts required for a reliable calibration. For these reasons, a research project (the "CALMO Priority Project") within the COSMO consortium has been defined and accepted. The project aims at conducting the basic research required for proving the effectiveness of the calibration framework developed at ETHZ in the context of regional climate simulations for NWP applications, making the necessary adaptations, and assessing its practicability.

The proposed objective calibration methodology has the potential to bring a transformative change to atmospheric model development and significantly reduce model development cycle times. The calibration method will be a very useful tool to improve the quality of the multiple configurations of atmospheric models running in Europe and beyond. More specifically, the developed methodology could be used by each COSMO member to define an optimal calibration over the target area of interest, for re-calibration after major model changes (e.g. higher horizontal and / or vertical resolution), as well as for an unbiased assessment of different modules (e.g. parameterization schemes), and for optimal perturbation of parameters when run in ensemble mode. Furthermore, a better understanding of the sensitivity of the model quality associated with a specific parameter value could benefit the quantification of the flow dependent model forecast error. Last but not least, the implementation of the methodology for a specific parameter can clarify the impact of the specific parameter on the overall model performance. Once the meta-model has been fitted to the full COSMO NWP model both the effect of the parameter setting and parameter space used (i.e., the maximal range of optimal values) can be determined without the use of the full NWP model.

Scientific goals and objectives

The main scientific goals and objectives are in-line with the CALMO research project, which has been proposed within the COSMO Consortium and accepted in September 2012. Namely, they are:

- **Provide an objective methodology for NWP models** that can substitute expert tuning: Establish a standard procedure (tool) that objectively improves NWP model performance by optimally determining unconfined parameters.
- **Understand the sensitivity of the NWP model quality** with respect to the model parameter space.
- **Optimize the calibration procedure** with respect to the required amount of computing resources for each re-calibration.

As a demonstration vehicle, the main COSMO configuration to be calibrated in this project is the 1.1 km mesh-size COSMO model version currently developed within the COSMO-NExT [1] project at MeteoSwiss, to be operational in 2016. For development of the calibration method, we will use a 2.2 km version of the model, which is computationally much less expensive. This will yield an objective inter-comparison between the known 'expert tuning' of that version as well as an inter-comparison between two model version.

Significant experience has already been gathered since the start of the CALMO project. Currently, the adaptation of the methodology is being investigated over a large domain (covering Europe) and with a 7 km mesh-size. Due to computational constraints, initial tests are performed at these coarse mesh-sizes and using only two parameters for calibration (namely the laminar boundary layer roughness and the minimal diffusion coefficient for heat) for the entire year 2008. Preliminary results show a low sensitivity on both 2 m temperature and precipitation. The reasons for this are not yet fully understood but may indicate that the selected parameters are not the most sensitive ones concerning NWP forecast quality. The current implementation of the method should urgently be shifted to mesh-sizes which are more relevant to current and future model implementations and expanded to more than two parameters. As a consequence, the number of required simulations will increase and a significant amount of computing resources are required. The selection of additional parameters to be calibrated is done in view of variables that are essential to weather forecasting, mainly 2 m temperature and precipitation. Initial emphasis is given on selecting parameters affecting different physical processes, ranging from cloud-radiation interaction to microphysics and boundary layer processes.

An important aspect to be considered for the implementation of the method is the selection of the performance score for the model quality since it critically influences the sensitivity of the forecast quality with respect to the various poorly defined model parameters. For temperature, we intend to use root mean squared error (RMSE) as the performance score, which is widely used for temperature verification (Murphy, 1988). For precipitation it is known (Katz and Murthy, 1997) that several different accuracy measures have to be used to fully assess the value of the forecast. Stable Equitable Error in Probability Space (SEEPS) proposed by Rodwell et al. (2010) as well as FBI, ETS, and TSS proposed in the work of Cherubini et al. (2002) can be used. Special attention is required when selecting the observations used for the determination of the model performance, as they need to be consistent with the modeling framework and resolution. High-quality gridded datasets of both precipitation and temperature are available for long time periods (e.g. Isotta et al., 2013; Frei, 2013) and will be used for the verification of the calibrated model. Once the simulations have been performed, different scores or combinations thereof can be applied rapidly and without requiring further model integrations.

The third important objective of this project is to optimize the calibration procedure with respect to the required amount of computing resources for each re-calibration: the less resources are needed for a calibration, the more (re-) calibrations can be done. While today expert tuning is often only

[1] http://www.meteoschweiz.admin.ch/web/en/research/current_projects/forecast/COSMO-NExT.html

done once for a specific model parameter, an objective and reproducible calibration should ideally be run as often as needed (i.e., after every major model change, for an unbiased comparison of different model formulations, to avoid or remove compensation errors, etc.). It is currently not a priori clear what the best minimal simulation strategy is for a robust objective calibration. In case of COSMO-CLM, Bellprat et al. (2012a) use 5 years of monthly means (resulting in 60 data points per parameter) to determine the meta-model coefficients. It is unlikely that the same statistics (which would mean a six-fold reduction of the required computing resources as compared to the needs to simulate a full year) will also be sufficient to calibrate an NWP model since daily means (e.g., 24h precipitation sums) exhibit a larger variability than monthly means, but it will be an important finding of this project to determine the minimal required data-set to perform a (regular) objective NWP model calibration. We plan to investigate different thinning strategies (only every *n*-th forecast, only *n* weeks every season, etc.) and their impact on constraining the optimal values of calibration parameters. Additionally, we will investigate the usage of coarser resolution simulations (2.2 km) for the calibration of more expensive high-resolution simulations (1.1 km). As a consequence, we require a considerable amount of computing resources in order to generate a robust reference calibration with a full year of 1.1 km daily forecasts.

Description of the research methods, algorithms, and code

In this section we briefly introduce the code that we will use for the simulations and describe the experimental design for applying the objective calibration method described above.

The code used in this project is the limited area numerical weather prediction (NWP) model COSMO (http://www.cosmo-model.org/). COSMO employs finite differences to solve the primitive hydro-thermodynamic equations that describe the non-hydrostatic compressible flow of the atmospheric constituents. It contains a comprehensive set of parameterization schemes to represent non-resolved processes. COSMO is suitable for running simulations with grid spacings ranging from O(100km) to O(1km). The COSMO model has been running for many years at CSCS, both for numerical weather prediction and regional climate research, on several generations of CSCS supercomputers.

Within the HP2C (http://www.hp2c.ch/) project "Regional climate and weather modeling on the next generations high-performance computers" the COSMO model has been the main target application. The dynamical core was restructured to be more easily adapted to new emerging computer architectures. The dynamical core is now written in C++ and is based on the stencil library STELLA (Gysi et al., 2014) and on a new communication library developed at CSCS (Bianco, 2013). In the HP2C project "Operational COSMO Demonstrator (OPCODE)" other important parts like the necessary physical parameterization schemes were ported to GPUs mainly through the use of OpenACC compiler directives in the respective Fortran code (Lapillonne and Fuhrer, 2013). For this project, this refactored version of the COSMO model capable of running on GPU-based hardware architectures will be used (referred to as RC). By the time of the start of this project, the RC version will be available based on the version 5.0 of COSMO. The benchmarks provided below, are done using the RC version based on COSMO 4.19.

The computationally most expensive model setup will be based on the research configuration MeteoSwiss is using for the 1.1 km model named COSMO-1, currently under development. Such a high horizontal resolution is unprecedented and is currently not used for numerical weather prediction. The domain has a size of 1158 x 774 grid points in the horizontal and 80 vertical levels and spans the greater Alpine region. The simulation domain and the associated topography (in meters above sea-level) are shown in Fig. 3. The 2.2 km simulations will be executed over the same domain but have only 60 vertical levels, and thus the computational cost is reduced by a factor of approximately 10 with respect to a corresponding 1.1 km simulations.

The calibration will be performed using 8 parameters. As already stated, the calibration requires at least 2*N + N * (N-1) / 2 repetitions of a single model experiment where N is the number of

parameters used, each experiment being characterized by a different set of model parameters. As a first step, the computationally much less demanding COSMO-2 will be calibrated (with 6 or 7 parameters to be calibrated) to gain experience in the calibration methodology and its applicability to NWP models. Then, one year long COSMO-1 simulations in forecast mode (i.e. no direct use of observations) will be performed. The year 2008 has been chosen, because it is representative for a mean climatology over Europe.

The initial and boundary conditions for all experiments will be taken from an analysis run at 2.2 km resolution. An important issue that requires careful consideration is the initialization of the soil, since the typical time-scale for the adjustment of the soil moisture to a change in the model climate is of the order of years. One possible approach will be a spin-up run with a much cheaper coarser model, before starting each calibration experiment.

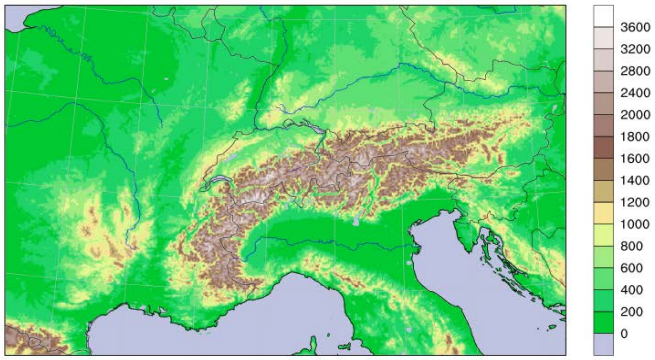

Figure 3: Orography and area covered by the 1158 x 774 grid point domain.

Parallelization approach

The standard COSMO version employs a two-dimensional domain decomposition along the two horizontal directions with a pure MPI parallelization strategy. Inter-node communication is dominated by nearest neighbor communication between adjoining sub-domains in order to update ghost values required for the application of the finite difference operators, which is implemented using a MPI_ISend and MPI_Recv scheme. Communication between accelerators is done using G2G technology avoiding transfers from device to host and vice versa. The G2G transfers are implemented using a MPI_IRecv/MPI_ISend/MPI_Wait strategy and overlap computation with communication wherever possible. The only remaining collective communication patterns are either for computing scalar diagnostics or in the I/O part of the code in order to gather fields for writing to the storage subsystem. Fine-grain, on-node parallelization is done using CUDA threads in the horizontal dimensions, where a minimum of approximately 96 x 96 gridpoints per accelerator are required for efficient execution (see below).

Since the research in this project requires an ensemble of simulations with different tuning parameters, the ensemble members are completely independent and can be run in parallel. The number of nodes required by each ensemble member is determined by the scalability of the model and the required time-to-solution for the ensemble simulation.

Representative benchmarks and scaling

In this section we present two types of performance benchmarks. First, we present single socket scaling results in order to assess the typical per-node problem sizes that can efficiently be offloaded to GPU-based hardware. Second, we present benchmarking results from the Cray XC30 (Piz Daint) currently open for production at CSCS using a representative COSMO simulation at 1.1 km horizontal resolution. Piz Daint is hybrid high performance computing system with one Intel Xeon E5-2670 and one NVIDIA Tesla K20X per node. We will use the benchmark results for our estimation of required resources (see below).

Single socket scaling

The single socket scaling behavior of the refactored COSMO code in the absence of output operations is shown in Figure 4 for a node with two sockets of Intel Xeon E5-2670 as well as a node with a single NVIDIA Tesla K20x. The x-axis shows the number of horizontal grid points (atmospheric columns) that are assigned to one node. The results shown here exclude inter-node communication overheads, since only single node simulations have been executed. On both processor types there is an almost linear dependence on the number of grid points, until the number of grid points falls below a certain threshold. For GPUs this threshold is reached at approximately 96 x 96 grid points per node, whereas for CPUs scaling continues down to 16 x 16 grid points per node. For efficient GPU execution, a minimum number of grid points per node is required. Hence, there is a clear limit on the number of GPUs that can be efficiently used in parallel for a given problem size as well as on the minimum attainable time-to-solution. As long as we can stay within the range where GPUs are operating efficiently the use of GPUs is favorable.

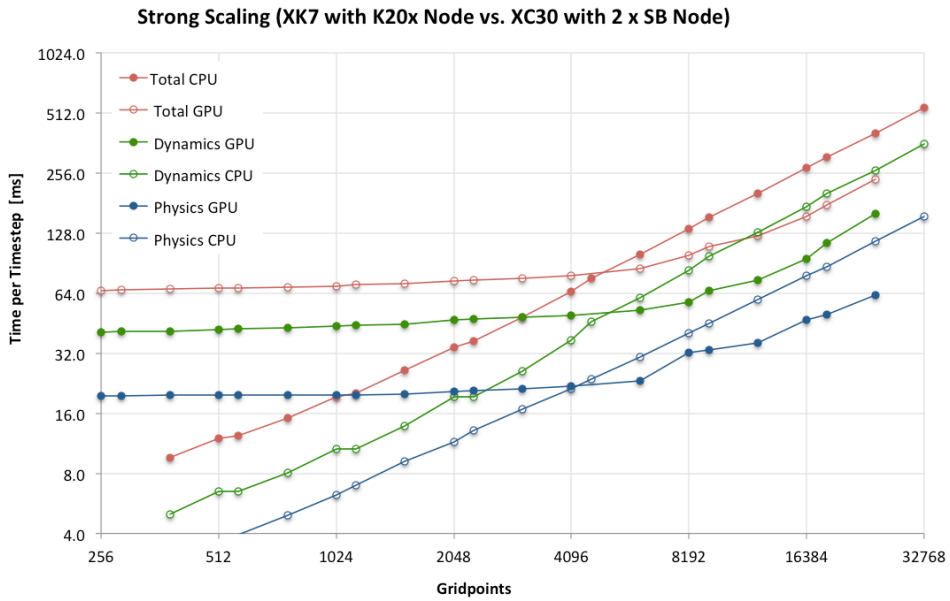

Figure 4: Wall clock time needed for a single time step with the restructured COSMO model on a single socket as a function of the number of grid points assigned to this socket for Nvidia Tesla X2090 GPU (black), Intel SandyBridge CPU (solid circles) and AMD Interlagos CPU (hollow circles)

Full benchmarks

Table 1 shows the benchmark results on Piz Daint hybrid nodes for the restructured code (RC) using the NVIDIA Tesla K20x accelerators (GPU) and the Intel Xeon E5-2670 processors (CPU) for a simulation representative for one ensemble member with a forecast time of +24 hours including I/O. The domain size and model configuration are representative for the runs we plan to execute except for using version 4.19 of the code. Initial testing has shown, that moving from version 4.19 to 5.0 will increase the total runtime of the model by approximately +15%, due to improvements of the numerical accuracy in the fast waves solver. Table 1 shows that the CPU version shows good strong scalability retaining a parallel efficiency of 85% when comparing a run on 16x16 nodes against a 8x8 node reference run. The GPU version shows poor strong scalability. Memory constraints on the GPU do not allow us to put more than approximately 128x128 grid points on a single GPU, and thus we are not able to decrease the number of nodes below 8 x 8 for this problem size.

Figure 5 shows the scaling behavior of the restructured version (RC) in CPU and GPU mode measured on Piz Daint. This experiment takes into account also the effects of inter-node communication and output to mass storage. This explains the differences to Table 1, which did not contain any inter-node communication and I/O.

Configuration	Version	# nodes	#GP/ node	Runtime [min]	nodeh/modelyr	speedup	efficiency
RC 8x8	**GPU**	**65**	**13824**	**47.2**	**18665**	**REF**	**REF**
RC 10x10	GPU	101	8847	45.4	**27884**	1.0	0.67
RC 12x12	GPU	145	6144	42.3	**37312**	1.1	0.50
RC 14x14	GPU	197	4514	38.3	**45899**	1.2	0.40
RC 16x16	GPU	257	3456	42.2	**65976**	1.1	0.28
RC 8x8	CPU	65	13824	104.3	**41217**	REF	REF
RC 10x10	CPU	101	8847	74.4	**45712**	1.4	0.90
RC 12x12	CPU	145	6144	51.2	**45150**	2.0	0.91
RC 14x14	CPU	197	4514	39.6	**47469**	2.6	0.86
RC 16x16	CPU	257	3456	30.8	**48205**	3.4	0.85

Table 1: Benchmark results on Piz Daint (Cray XC30) for a +24 hour simulation with the target domain size of 1158 x 774 grid points at 1.1 km resolution. Results of experiments for the restructured code (RC) running on accelerators.

For the production simulations of this project the ratio between wall clock time and model time is of key importance, as it determines the time-to-solution for a given simulation period. In order to complete a 1 year long simulation within approximately 1 to 2 months a ratio of at least 12 must be attained (accounting for queue wait time and failure recovery). Thus a 24 hour simulation should have a wall clock time of approximately 120 min or less. Table 1 shows that in terms of nodeh/modelyr the GPU version of the code is over a factor 2.2 more efficient. **For the configuration using 8 x 8 + 1 = 65 nodes and running on GPU, we require a total of 12 wall-clock days for a simulation of 1 model year. This is a sufficient time-to-solution for this project and we plan to use this setup for the production runs.** Since the simulation plan contains an ensemble simulation with completely independent ensemble members, we expect to be able to simulate the full ensemble within three to four months of wall-clock time, accounting for system usage by other users, queue wait time and failure recovery.

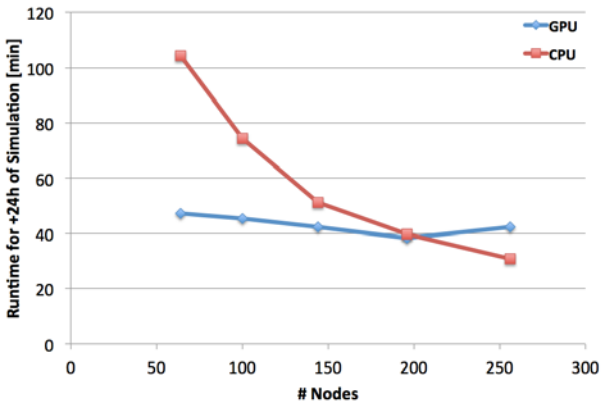

Figure 5: Wall clock time [min] for +24 hours of model time for a domain of size 1158 x 774 gridpoints as a function of numbers of nodes used for the simulation for the GPU version (blue) and the CPU version (red) of COSMO measured on Piz Daint.

Performance analysis

Table 2 below shows the required performance analysis metrics for the benchmark. The benchmark corresponds to a COSMO configuration at 1.1 km resolution integrated forward 24 hours with activated performance profiling using CrayPat (pat_build –g mpi,io). The results were obtained on Piz Daint with the restructured code on CPU and GPU for the target configuration using 8x8 nodes (see above). It should be noted, that some of these performance metrics only make limited sense in the case of GPU execution, but we have included them to comply with CSCS' requirements for production proposals.

Architecture	CPU	GPU
Number of nodes	64 + 1	64 + 1

Number of MPI ranks	512 + 1	64 + 1
Wallclock time [s]	7403.5	3605.1
Memory [MB] / process	710.8	3163.1
MPI (% of total walltime)	11.5%	8.8%
MPI_SYNC (% of total walltime)	6.5%	3.7%
MPI call1 (% of total walltime)	MPI_Wait 6.7%	MPI_Recv 7.4%
MPI call2 (% of total walltime)	MPI_Allreduce 5.5%	MPI_Allreduce 1.5%
%peak (DP)	12.3%	3.3% (not applicable)
PAPI FP OPS / process	2.13E+13	2.87E+12
PAPI L1 DCM / process	2.82E+13	9.64E+12
Write Rate (MB; MB/sec)	66594.6 MB 100.2 MB/s	64602.8 MB 743.14 MB/s

Table 2: Selected output values from CrayPat for the benchmarks running on 8x8 nodes with the 1.1 km resolution configuration of the restructured COSMO code on Piz Daint

As is typical for weather applications, applying finite difference discretizations for the solution of the underlying governing equations, per process memory usage is very low. MPI communication is dominated by nearest neighbor communication (MPI_Recv) and synchronization. The comparatively large synchronization times stem from inherent imbalances in the model, mostly caused by modules of differing cost being applied to sea/land grid points and/or cloudy/cloud free regions. The floating point efficiency depends on several factors (local domain size as compared to cache size, vectorization along the first horizontal direction, etc.) and is not straightforward to explain.

Project plan: tasks and milestones

This project will be a one year project running from 1.10.2014 to 30.9.2015. The distribution of individual tasks over the project period is shown in Table 3. During the first three months a calibration of the 2.2 km mesh-size COSMO version will be performed using 6 to 7 parameters to test the calibration method for a convection-permitting COSMO configuration. In computational terms the calibration at 2.2 km is a factor 10 cheaper than the one for 1.1 km horizontal resolution, and it is hence well suited as a start-up exercise. Moreover, it will also allow for an objective inter-comparison between the two model versions, and provide an assessment of the added value of higher resolution.

Months 4 to 12 are devoted to one year simulations (one forecast per day) with 1.1 km resolution and 8 perturbed parameters. This requires substantial computing resources (990'000 node hours, see below). This proposal requests to get these resources in full, in order to generate a complete data set and a robust reference calibration. For a calibration tool that needs to be run as often as possible (ideally after every major change of the model configuration, after major code changes or after implementation of new model components such as new parameterization schemes, etc.) this large amount of computing resources is a major stumbling block. We will therefore thoroughly investigate to what extend the data set of full model runs can be reduced to still obtain a robust and good quality calibration result.

Task	Description	Months	Milestones
1	Refinement and testing of the calibration method with COSMO-2.	1-3	COSMO-2 simulations are done, performance score for NWP is defined, model parameters to be calibrated is determined.
2	Full year simulations with COSMO-1. Parameter	4-12	COSMO-1 simulations are done (mainly in months 4-9), meta-model is set up, minimal

	estimation for coefficient of meta-model using the full one-year data set as well as sub-sets thereof.		amount of model runs needed for a robust calibration is determined, COSMO-1 is calibrated.

Table 3: Distribution of tasks and milestones over the year.

We expect a high interest from the broader research community in the findings of this project and plan to publish the results in a high-profile, peer-reviewed journal.

The project team which will execute the proposed research and simulations has the required skills to perform the tasks in a professional, effective and timely manner. Dr. Voudouri is the project leader of the CALMO project and will carry the overall responsibility. The Co-PI's (Dr. Bettems and Dr. Khain) are both CALMO project members. Dr. Bettems has significant experience in running production simulation workflows for NWP on the CSCS systems. The team will profit from existing scientific collaboration with Prof. Schär and Dr. Bellprat, who have developed the methodology for climate applications. The GPU version of the COSMO model is being developed by a team under the lead of Dr. Fuhrer, who has significant experience with running COSMO on hybrid Piz Daint. Finally, Dr. Arpagaus is the project leader of the COSMO-NExT project, developing the next generation weather prediction models for Switzerland. He adds significant experience in model validation and verification and will ensure the applicability of the methodology to a real world NWP system.

Resource justification

The following Table 4 shows the resources needed for the project. The calculations are based on the benchmark and scaling tests given in the two preceding sections. The configuration chosen for the needed resource estimates is RC 8x8 from Table 1 above. Due to the upgrade of the code version from 4.19 to 5.0 we expect a further increase in runtime of approximately +15%, due to the introduction of a new fast-waves solver in the dynamical core. For the estimation of needed resources, we added a margin of 5% to the benchmark result in Table 1 to account for jobs that need to be re-run because of failures of any kind (system problems, data problems etc.). In summary, we thus require 22'500 nodeh/year for the RC 8x8 configuration (at 1.1 km resolution) on GPU.

Simulation type	**Number of simulations**	**nodeh/year**	**Total nodeh**	**Storage needs**
One year simulations with 2.2 km resolution for 6 to 7 parameters	35 x 1 year (GPU)	**2'250**	**78'750**	**12 TB**
One year simulations with 1.1 km resolution for 8 parameters	44 x 1 year (GPU)	**22'500**	**990'000**	**58 TB**
Total			**1'068'750**	**70 TB**

Table 4: Overview of needs for the allocation period

The amount of raw data produced with the simulations will be of the order of 70 TB for the entire amount of simulations. The actual amount of storage capacity needed is extremely high as we intend to perform simulation that cover in total 44 years. To address this problem, a data thinning policy will be employed. Only 2-dimensional fields for specific variables will be kept for the entire simulation period. Thus, the amount of storage space needed is significantly reduced. The numbers given in the table above already take into account this reduction especially for the 44 one year simulations with the 1.1 km configuration.

References

Annan, J. D., and J. C. Hargreaves. 2007. Efficient estimation and ensemble generation in climate modelling. *Philosophical Transactions of the Royal Society A-mathematical Physical and Engineering Sciences,* **365**, 2077-2088.

Bayler, Gail M.; Aune, R. M.; Raymond, W. H., 2000. NWP Cloud Initialization Using GOES Sounder Data and Improved Modeling of Nonprecipitating Clouds. *Mon. Wea. Rev.*, **128**, 3911-3921.

Bellprat, O., S. Kotlarski, D. Lüthi, and C. Schär. 2012a. Objective calibration of regional climate models. *Journal of Geophysical Research,* **117**, D23115.

Bellprat, O., S. Kotlarski, D. Lüthi, and C. Schär. 2012b. Exploring perturbed physics ensembles in a regional climate model. *Journal of Climate,* **25**, 4582-4599.

Bellprat, O., S. Kotlarski, D. Lüthi, C. Schär, A. Frigon, R. De El__a, and R. Laprise. n.d. Spatial transferability of objectively calibrated regional climate models. *In preparation.*

Bianco M. 2013. An interface for halo exchange pattern. http://www.prace-i.eu/IMG/pdf/wp86.pdf.

Cherubini T., Ghelli A., Lalaurette F. ,2002. Verification of Precipitation Forecasts over the Alpine Region Using a High-Density Observing Network. *Mon. Wea. Rev.*, **17**,238-249.

Duan Q., J. Schaake, V. Andre´assian, S. Franks, G. Goteti, H.V. Gupta, Y.M. Gusev, F. Habets, A. Hall, L. Hay, T. Hogue, M. Huang,G. Leavesley, X. Liang, O. Nasonova, J. Noilhan, L. Oudin,S. Sorooshian, T. Wagener, E.F. Wood. 2006. Model Parameter Estimation Experiment (MOPEX):An overview of science strategy and major results from the second and third workshops. *Journal of Hydrology*, **320**, 3–17.

Frei, C. 2013: Interpolation of temperature in a mountainous region using nonlinear profiles and non-Euclidean distances. *Int. J. Climatol.* Published online: DOI:10.1002/joc.3786

Gysi, T., O. Fuhrer, C. Osuna, T. Schulthess, 2014: STELLA: A DSEL library for performance portable implementation of stencil computations on structured grids. *In preparation.*

Isotta, F. A. Frei C, Weilguni V, Perčec Tadić M., Lassègues P., Rudolf B., Pavan V., Cacciamani C., Antolini G., Ratto S., Munari M., Micheletti S, Bonati V., Lussana C., Christian Ronchi C., Panettieri E., Marigo G and Vertačnik G, 2013: The climate of daily precipitation in the Alps: development and analysis of a high-resolution grid dataset from pan-Alpine rain-gauge data. *Int. J. Climatol.,* Published online, DOI: 10.1002/joc.3794

Katz W. R. and A.H. Murthy, 1997. Economic value of weather and climate forecasts. Cambridge University Press, 222 pp.

Knutti, R., T. F. Stocker, F. Joos, and G. K. Plattner, 2002: Constraints on radiative forcing and future climate change from observations and climate model ensembles. *Nature*, **416 (6882)**, 719-723.

Kotlarski, S. K. Keuler, O.B. Christensen, A. Colette, M. Déqué, A. Gobiet, K. Goergen, D. Jacob, D. Lüthi, E. van Meijgaard, G. Nikulin, C. Schär, M. Suklitsch, C. Teichmann, R. Vautard, K. Warrach-Sagi, V. Wulfmeyer, 2014: Regional climate modeling on European scales: A joint standard evaluation of the EURO-CORDEX RCM ensemble. *Geosci. Model Dev. Disc.*, **7**, 217-293.

Lapillonne, X., and O. Fuhrer, 2013: Using compiler directives to port large scientific applications to GPUs: An example from atmospheric science. *Parallel Processing Letters*, in press.

Myers R. H,Douglas C. Montgomery,Christine M. Anderson-Cook, 2009. Response Surface Methodology: Process and Product Optimization Using Designed Experiments. Wiley Series in Probability and Statistics.

Murphy, J., 1988: Skill scores based on the Mean Square Error and their relationships to the correlation coefficient. *Mon. Wea. Rev.,* **116**, 2417-2424.

Neelin, J. D., A. Bracco, H. Luo, J. C. McWilliams, and J. E. Meyerson. 2010. Considerations for parameter optimization and sensitivity in climate models. Proc. of the National Academy of Sciences of the United States of America, **107**, 21349-21354.

O'Hagan, A. (2006). Bayesian analysis of computer code outputs: a tutorial. *Reliability Engineering and System Safety*, **91**, 1290-1300.

Rodwell M., Richardson D., Hewson T. D.and Haiden T. 2010, A new equitable score suitable for verifying precipitation in numerical weather prediction. *Q. J. Royal Met. Soc*, **136**, 1344–1363.

Skamarock, W.C., 2004. Evaluating Mesoscale NWP Models Using Kinetic Energy Spectra. *Mon. Wea. Rev.*, **132**, 3019-3032.

Stephens, Graeme L., Si-Chee Tsay, Paul W. Stackhouse, Piotr J. Flatau, 1990: The Relevance of the Microphysical and Radiative Properties of Cirrus Clouds to Climate and Climatic Feedback. *J. Atmos. Sci.*, **47**, 1742–1754.

Webb M., Hugo Lambert F. and Gregory J., 2013. Origins of differences in climate sensitivity, forcing and feedback in climate models. *Climate Dynamics*, **40**, 677-707.