Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Home Affairs FDHA
Federal Office of Meteorology and Climatology MeteoSwiss

New operational applications at MeteoSwiss on a hybrid supercomputer

O. Fuhrer[1], M. Arpagaus[1], A. Walser[1], D. Leuenberger[1], X. Lapillonne[1], P. Spoerri[2], **P. Steiner[1]**

[1]Federal Institute of Meteorology and Climatology MeteoSwiss

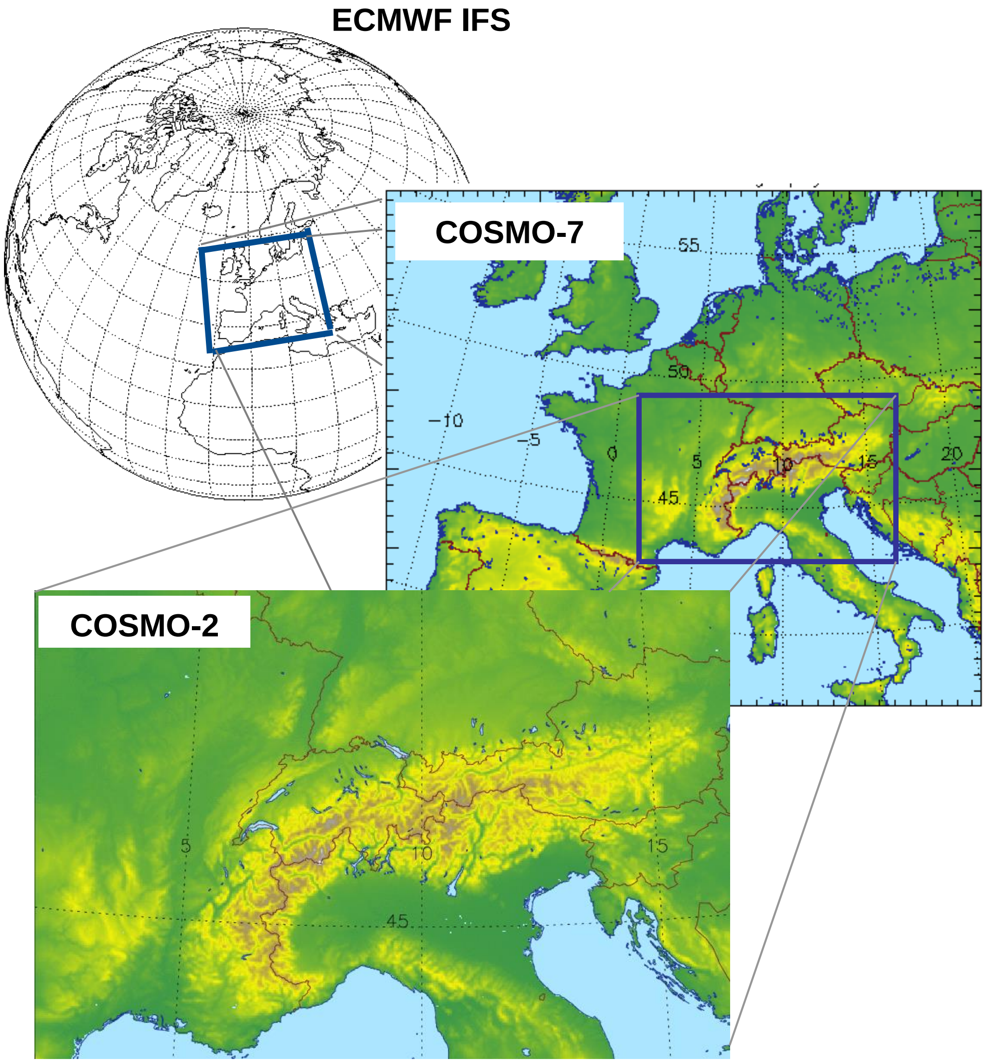
[2]Center for Climate Systems Modeling (C2SM), ETH Zurich

Current operational NWP system at MeteoSwiss

ECMWF IFS (global)
- 16km, 137 levels
- 2 x 240h per day

COSMO-7 (regional)
- 6.6km, 60 levels
- 3 x 72h per day

COSMO-2 (local)
- 2.2km, 60 level
- 8 x 33h per day



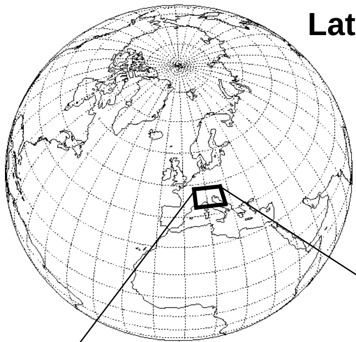
What the customers want …

- **better** (!) forecasts
- **higher resolved** (in space and time) forecasts
- ok, forecasts can not always be perfect, but then, please let me know ***when*** **the forecasts are bad, and how bad they are**
- **consistent** (in space and time) forecasts, i.e., across domain boundaries and lead-time limitations (… and forecasting systems!)
- **reliable** forecasts (quality as well as timeliness of delivery)

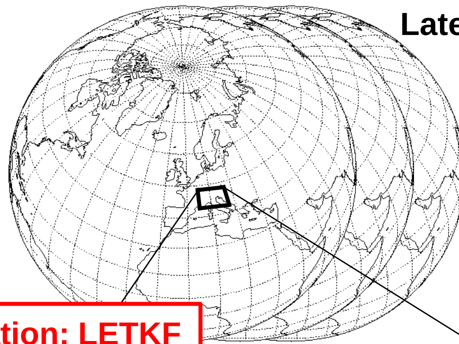
→ **Strategy of MeteoSwiss for its NWP system (2011)**

→ **Implementation in project COSMO-NExT (2012-2016)**

Future operational NWP system at MeteoSwiss



Lateral boundary conditions:
IFS-HRES
9km
4x per day



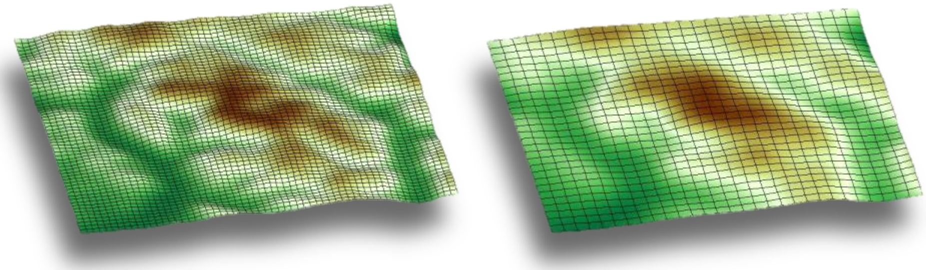
Lateral boundary conditions:
IFS-ENS
18km
4x per day

ensemble data assimilation: LETKF

COSMO-1: 33 hour forecasts, 8x per day
1.1km grid size (convection permitting)

COSMO-E: 5 day forecasts, 2x per day
2.2km grid size (convection permitting)
21 ensemble members

COSMO-1:
Setup vs. COSMO-2



- **Larger domain** (about 25%)
- **New code version**
- **More vertical levels** (80 instead of 60, using SLEVE)
- **No artificial horizontal diffusion** (except for flow dependent Smagorinsky type diffusion)
- **New upper boundary condition** (only vertical winds are being damped)
- **Higher frequency update of radiation** (every 6 minutes)
- **No parameterisation of sub-grid scale orographic drag**
- **No parameterisation of shallow convection**

→ **Skill is better than or equal to COSMO-2 in most parameters and seasons**

Philippe.Steiner@MeteoSwiss.ch

Summary Table COSMO-1 vs COSMO-2

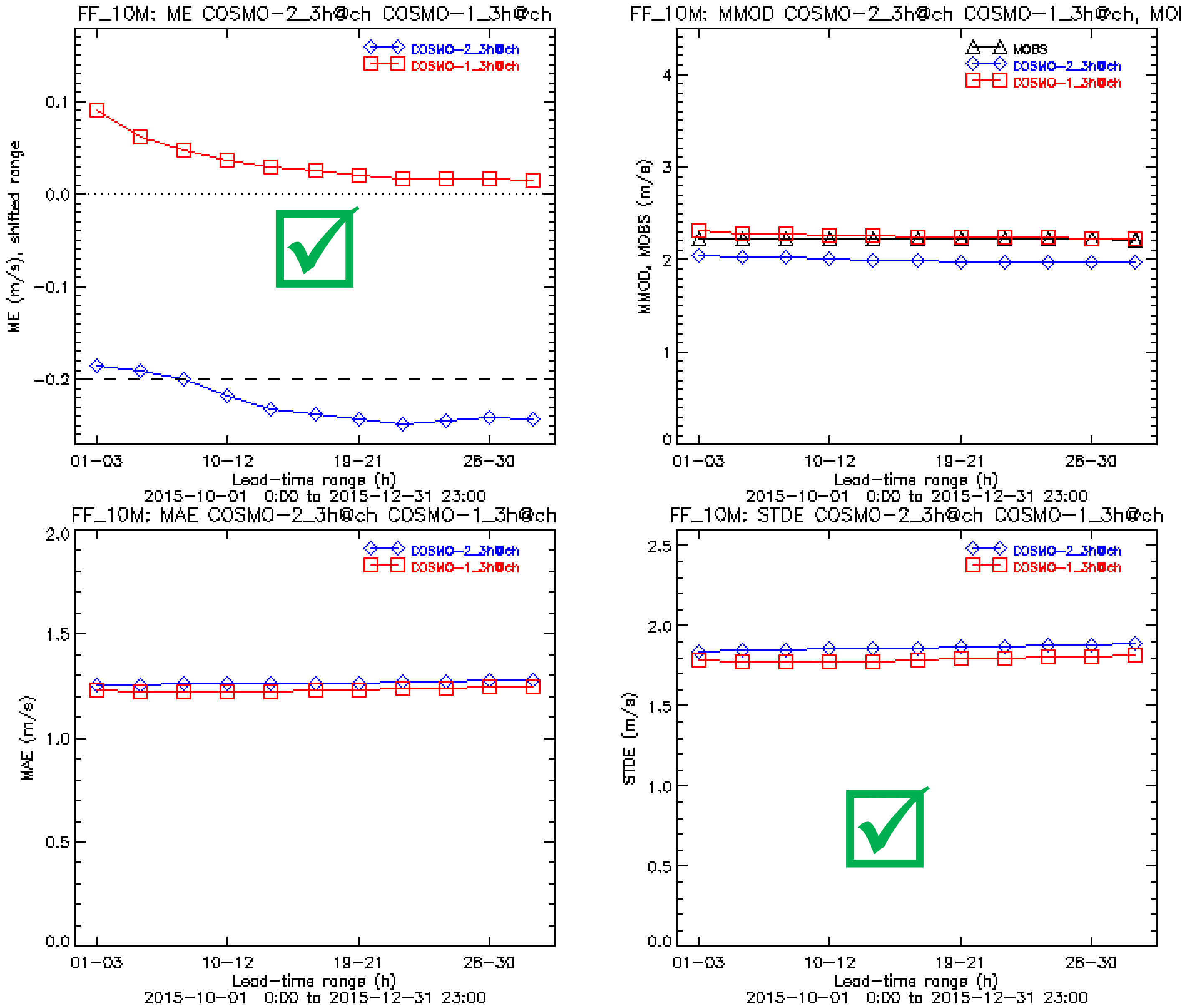
OND 2015 (total scores, all lead times, Swiss surface stations)

Parameter	ME	STDE	ETS Thrs 1	ETS Thrs 2	ETS Thrs 3
Surf. Pres.	☑|☑*	☑			
T 2m	☑	☑			
Td 2m	☑	☑			
dd 10m	☑*	☑			
ff 10m	☑	☑			
CLCT	☑	☑	☑	☑	
Prec 12h	☑	☑	☑	☑	☑
Prec 1h	☑	☑	☑	☑	☑*
Gusts	☑	☑	☑	☑	☑
Glob. Rad.	☑**	☑			

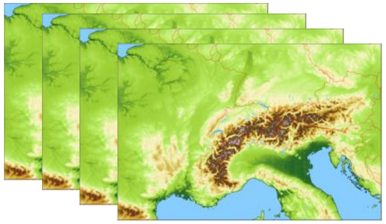
- ☑ Benchmark fulfilled
- * Slightly worse but insignificant
- ** Seemingly worse due to less compensating errors
- ⇧/⇩ Benchmark violated (over-/underpredicted)

Wind speed 10m COSMO-1 vs COSMO-2

OND 2015 (total scores, all lead times, Swiss surface stations)



COSMO-E



- Ensemble forecasts with convection-permitting resolution (2.2 km mesh-size) and 21 members
- Runs twice a day up to +120h for Alpine area
- Perturbations:
  - initial conditions: from LETKF
  - lateral boundary conditions: from IFS-ENS
  - model error: Stochastic Perturbation of Physical Tendencies (SPPT)
- Provides probabilistic forecast as well as “best estimate” of forecast uncertainty

→ **Skill is clearly better than COSMO-LEPS and at least as good as COSMO-2 in most parameters and seasons**

Summary Table COSMO-E vs COSMO-LEPS

SON 2015 (total prob. scores, all lead times, Swiss surface stations)

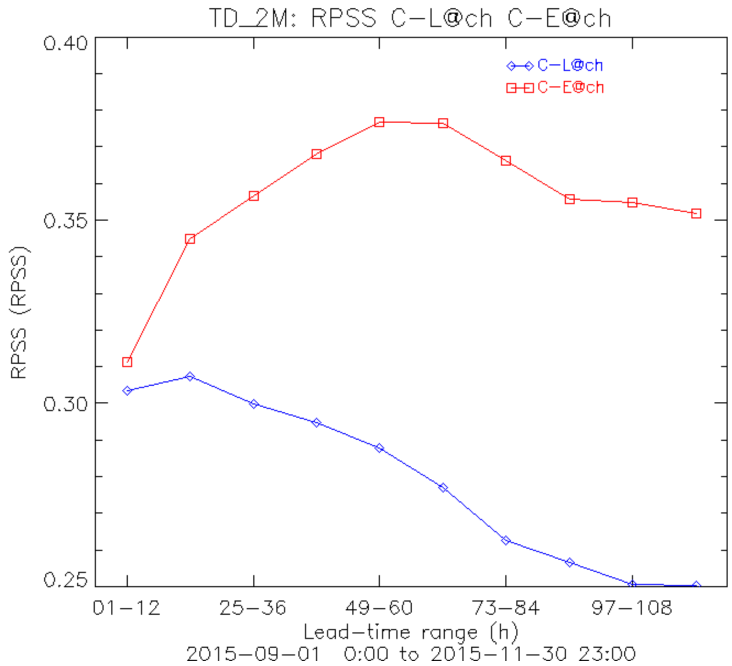
<table>
<tr><th>Parameter</th><th>RPS(S)</th><th>Outliers</th><th>Spread/ Error</th><th>Resolution Thrs1</th><th>Resolution Thrs2</th></tr>
<tr><td>T 2m</td><td>☑</td><td>☑</td><td>☑</td><td>☑</td><td></td></tr>
<tr><td>Td 2m</td><td>☑</td><td>☒|☑</td><td>☑</td><td>☒|☑</td><td></td></tr>
<tr><td>ff 10m</td><td>☑</td><td>☑</td><td>☑</td><td>☑</td><td></td></tr>
<tr><td>Prec 12h</td><td>☑</td><td>☑</td><td>☑</td><td>☒|☑</td><td>☑</td></tr>
<tr><td>Prec 1h</td><td>☑</td><td>☑</td><td>☑</td><td>☑</td><td>☒|☑</td></tr>
<tr><td>Gusts</td><td>☑</td><td>☑</td><td>☑</td><td>☑</td><td>☑</td></tr>
</table>

☑ Benchmark fulfilled
☒ Benchmark violated

Dew point 2m (SON 2015, Swiss surface stations)

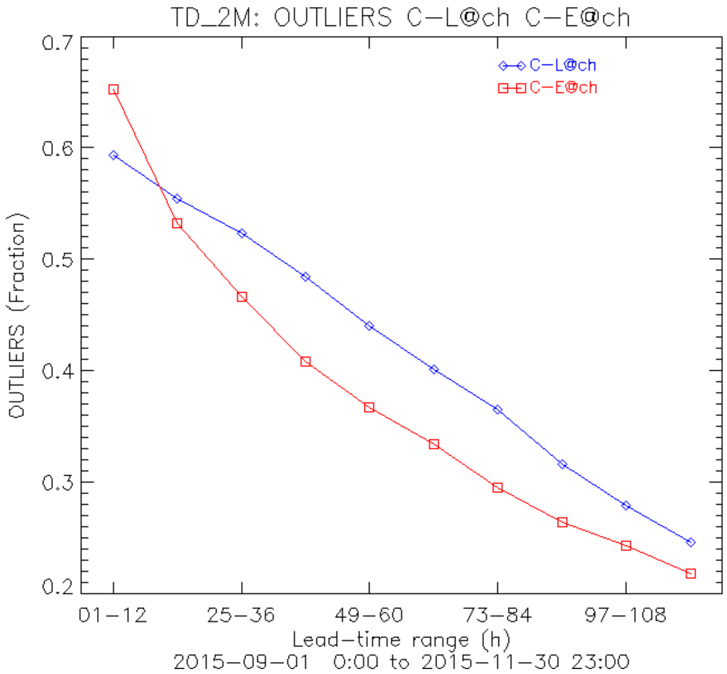
RPSS

COSMO-E
COSMO-LEPS



Outliers

COSMO-E
COSMO-LEPS

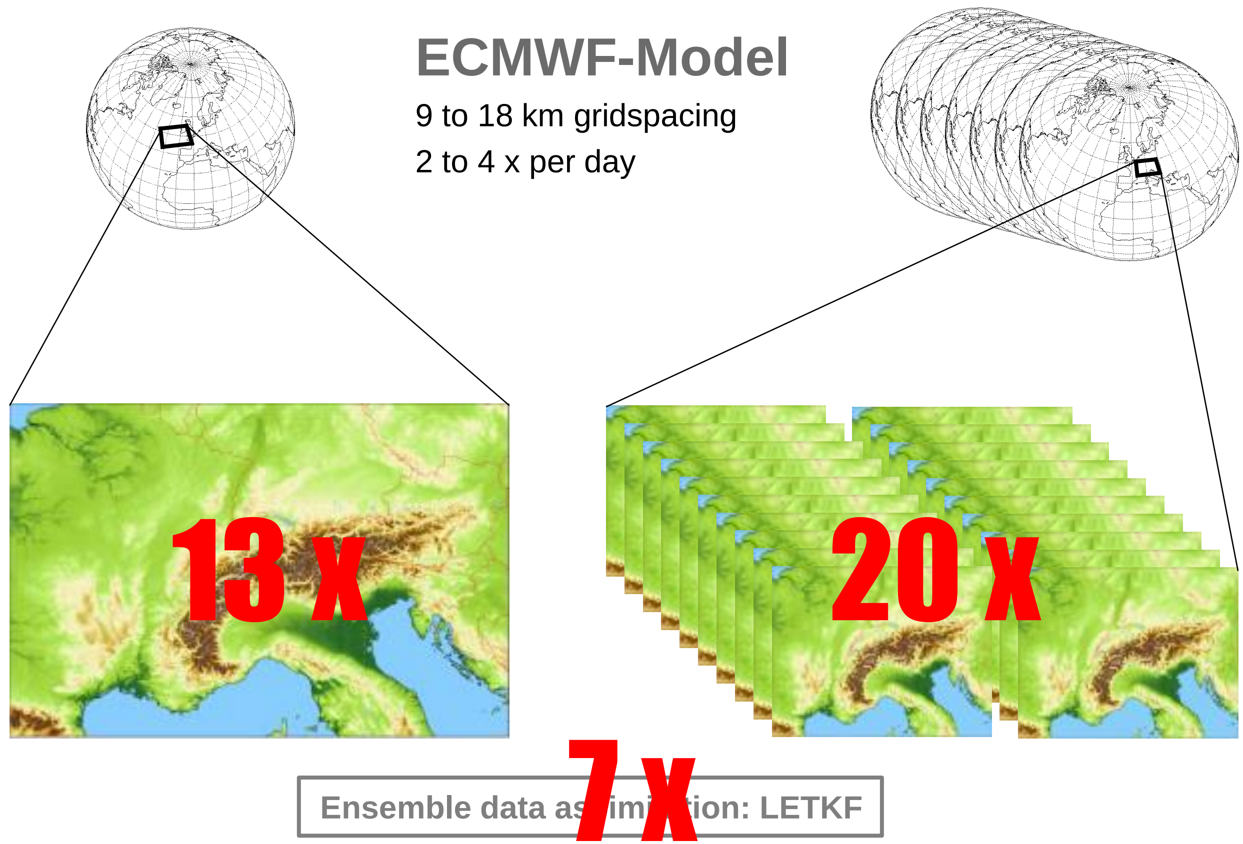


Computational cost = 40 x

(relative to current operational system)

COSMO-1

1.1 km grids size
8 x per day
1 to 2 d forecast



COSMO-E

2.2 km grid size
2 x per day
5 d forecast
21 members

Production with COSMO @ CSCS

Cray XE6 (Albis/Lema)

MeteoSwiss operational system

Since ~4 years

Next-generation system

Accounting for Moore's law (factor 4)

CSCS: Swiss National Supercomputing Centre (Lugano)

Images: CSCS

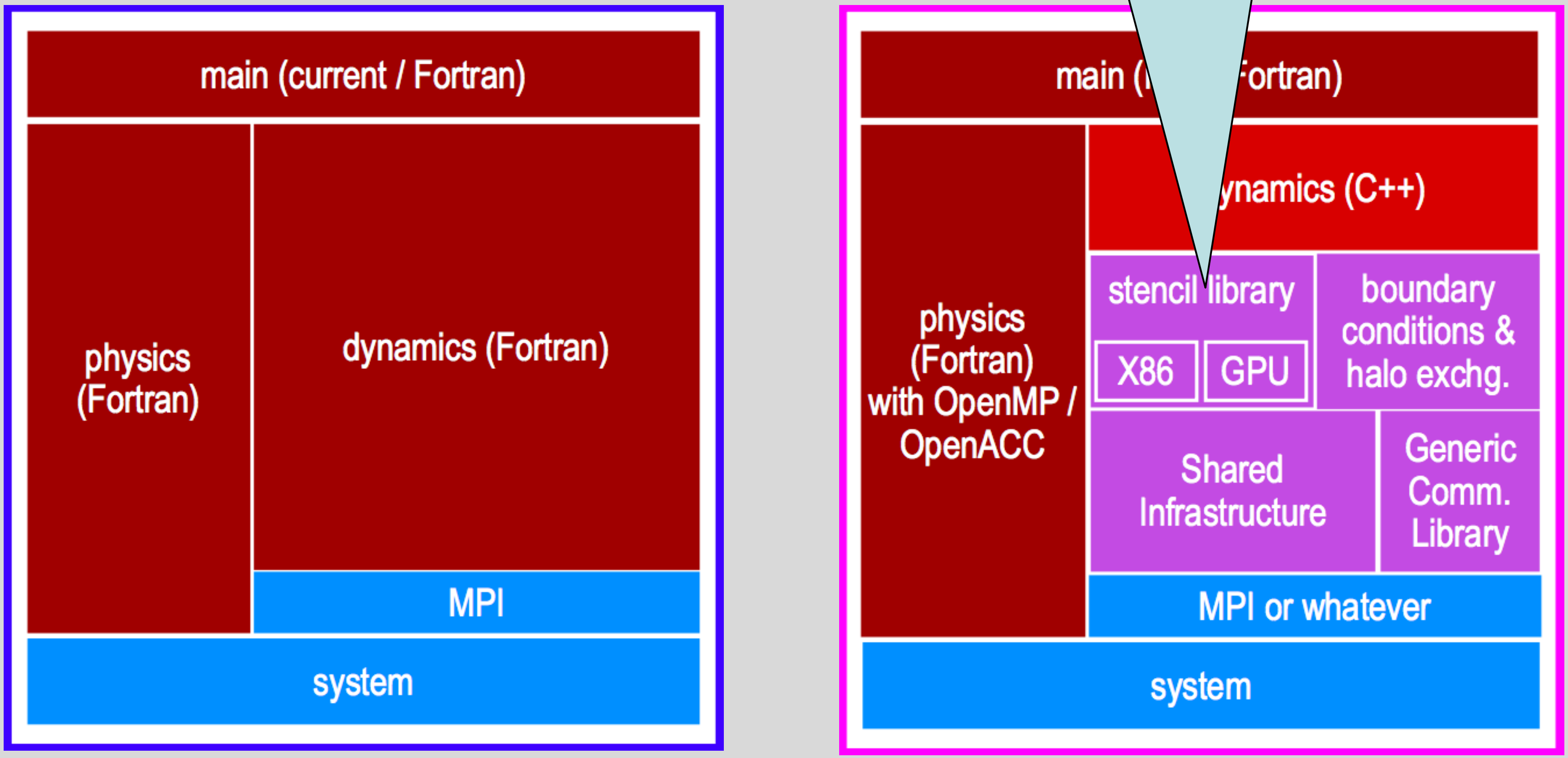
Choosen approach: co-design

- Design software, workflow and hardware with the following principles
  - Portability to other users (and hardware)
  - Achieve time-to-solution
  - Optimize energy (and space) requirements
- Collaborative effort between
  - MeteoSwiss, C2SM/ETH, CSCS for software since 2010
  - Cray and NVIDIA for new machine since 2013
  - Domain scientists and computer scientists
- Additional funding from Swiss HPCN Strategy (HP2C, PASC)

Images: CSCS

Current and new code

We are currently developing a more general version of STELLA: GridTools (global grids, FEM, …)

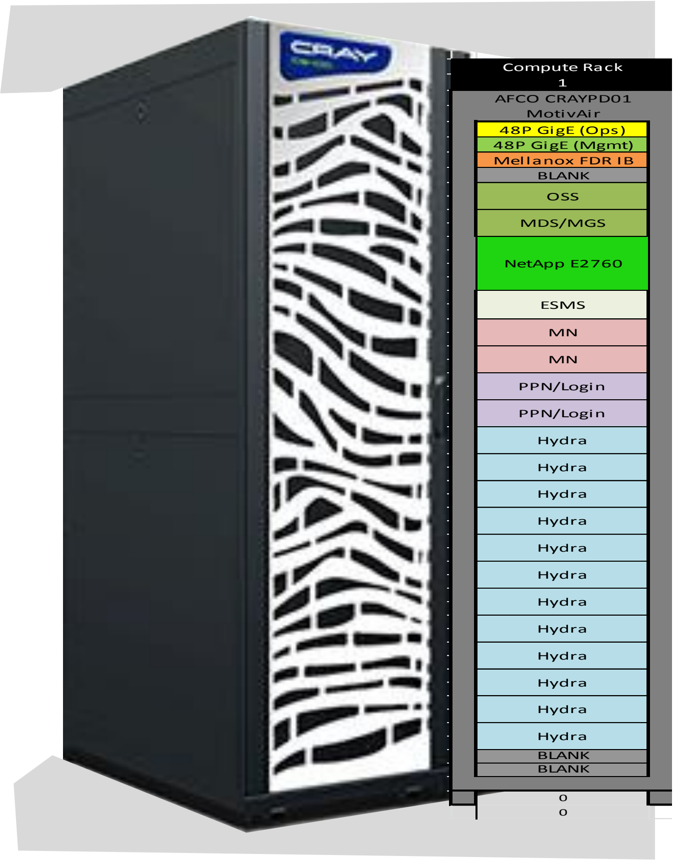


adapted from Fuhrer et al. 2014

New MeteoSwiss HPC system

Piz Kesch (Cray CS Storm)

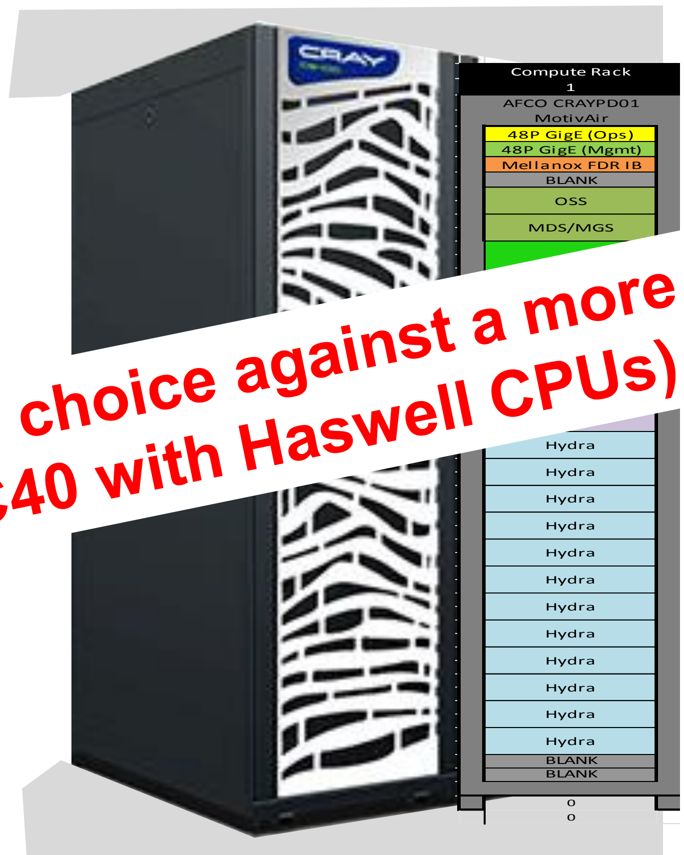
- Installed at CSCS in July 2015
- Hybrid system with a mixture of CPUs and GPUs
- “Fat” compute nodes with 2 Intel Xeon E5 2690 (Haswell) and 8 Tesla K80 (each with 2 GK210)
- Only 12 out of 22 possible compute nodes
- Fully redundant (failover for research and development)



New MeteoSwiss HPC system

Piz Kesch (Cray CS Storm)

- Installed at CSCS in July 2015
- Hybrid system with a mixture of CPUs and GPUs
- “Fat” compute nodes with 2 Intel Xeon E5 2690 (Haswell) and 8 Tesla K80 (each with 2 …
- Onl… …y redundant (failover for research and development)



It is now possible to compare our choice against a more “traditional” choice (e.g. Cray XC40 with Haswell CPUs)

CS Storm vs reference HPC system

Piz Kesch (Cray CS Storm)

- Installed at CSCS in July 2015
- Hybrid system with a mixture of CPUs and GPUs
- “Fat” compute nodes with 2 Intel Xeon E5 2690 (Haswell) and 8 Tesla K80 (each with 2 GK210)
- Only 12 out of 22 possible compute nodes
- Fully redundant (failover for research and development)

Piz Dora (Cray XC40)

- “Traditional” CPU based system
- Compute nodes with 2 Intel Xeon E5-2690 v3 (Haswell)
- Pure compute rack
- Rack has 192 compute nodes
- Very high density (supercomputing line)

Results
based on a COSMO-E benchn

Note: Not sure if this is an apples-to-apples comparison, due to different “character” of systems

	Piz Dora	Piz Kesch	Factor
Sockets @ required time-to-solution for 21 members	~16 CPUs	~7 GPUs	2.4 x
Energy per member	6.19 kWh	2.06 kWh	3.0 x
Time with 8 sockets per member	13550 s	5980 s	2.3 x
Cabinets required to run ensemble at required time-to-solution	0.87	0.39	2.2 x

Results relative to „old“ code

(„old“ = no C++ dycore, double precision)

	Piz Dora	Piz Kesch	Factor
Sockets at required time-to-solution for 21 members	~26 CPUs	~7 GPUs	3.7 x
Energy per member	10.0 kWh	2.06 kWh	4.8 x
Time with 8 sockets per member	23075 s	5980 s	3.8 x
Cabinets required to run ensemble at required time-to-solution	1.4	0.39	3.6 x

„Management summary“

Key ingredients

- Processor performance (Moore’s law) ~2.8 x
- Port to accelerators (GPUs) ~2.3 x
- Code improvement ~1.7 x
- Increase utilization of system ~2.8 x
- Increase in number of sockets ~1.3 x
- Target system architecture to application

Note Factor 4x comes from the software refactoring!

~ 40 x

Note Solution comes from a combination of investments in hardware, software and workflow

Image: Cray

Summary

- New forecasting system doubling resolution of deterministic forecast and introducing a convection permitting ensemble
- Co-design (simultaneous code, hardware and workflow re-design) allowed MeteoSwiss to increase computational load by 40x within 4–5 years
- Operations starting Q2 2016 on a CS Storm system with fat GPU nodes
- Energy to solution is a factor 3x smaller as compared to a "traditional" CPU-based system

References

O. Fuhrer, C. Osuna, X. Lapillonne, T. Gysi, B. Cumming, M. Bianco, A. Arteaga, T. C. Schulthess, “Towards a performance portable, architecture agnostic implementation strategy for weather and climate models”, Supercomputing Frontiers and Innovations, vol. 1, no. 1 (2014), see http://superfri.org/

G. Fourestey, B. Cumming, L. Gilly, and T. C. Schulthess, “First experience with validating and using the Cray power management database tool”, Proceedings of the Cray Users Group 2014 (CUG14) (see arxiv.org for reprint)

B. Cumming, G. Fourestey, T. Gysi, O. Fuhrer, M. Fatica, and T. C. Schulthess, “Application centric energy-efficiency study of distributed multi-core and hybrid CPU-GPU systems”, Proceedings of the International Conference on High-Performance Computing, Networking, Storage and Analysis, SC’14, New York, NY, USA (2014). ACM

T. Gysi, C. Osuna, O. Fuhrer, M. Bianco and T. C. Schulthess, “STELLA: A domain-specific tool for structure grid methods in weather and climate models”, to be published in Proceedings of the International Conference on High-Performance Computing, Networking, Storage and Analysis, SC’15, New York, NY, USA (2015). ACM