Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

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

Co-Designing a System for Regional Weather Prediction

Oliver Fuhrer[1], Xavier Lapillonne[1], Guilherme Peretti-Pezzi[2], Carlos Osuna[3], Thomas Schulthess[2]

[1]Federal Institute of Meteorology and Climatology MeteoSwiss
[2]Swiss National Supercomputing Centre CSCS, Lugano
[3]Centre for Climate Systems Modeling C2SM, ETH Zurich

.

Can you spot the weather model?

Data set:
11.8.2014 00.00:00 UTC - 13.8.2014 23.57:30 UTC

Spatial resolution (temporal resolution):
Model: 1.1 km (2.5'), satellite: approx. 4-5km (5'), weather radar: approx. 1km (5')

Simulations:
Dr. Oliver Fuhrer (MeteoSwiss)

Model computing:
Cray XK7 (GPU/CPU hybrid supercomputing system), Swiss National Supercomputing Centre CSCS

Visualization, data processing:
Dr. Oliver Stebler, Dr. Urs Beyerle

Background data:
Reto Stöckli, NASA Earth Observatory (NASA Goddard Space Flight Center)

Visualization project lead:
Dr. Oliver Stebler (oliver.stebler@env.ethz.ch)

Supervision:
Prof. Dr. Reto Knutti

Version:
23.1.2015

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

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

C2SM
Center for Climate
Systems Modeling

Current operational system

ECMWF-Model

16 km gridspacing

2 x per day 10 day forecast

COSMO-7

Δx = 6.6 km, Δt = 60 s

393 x 338 x 60 cells

3 x per day 72 h forecast

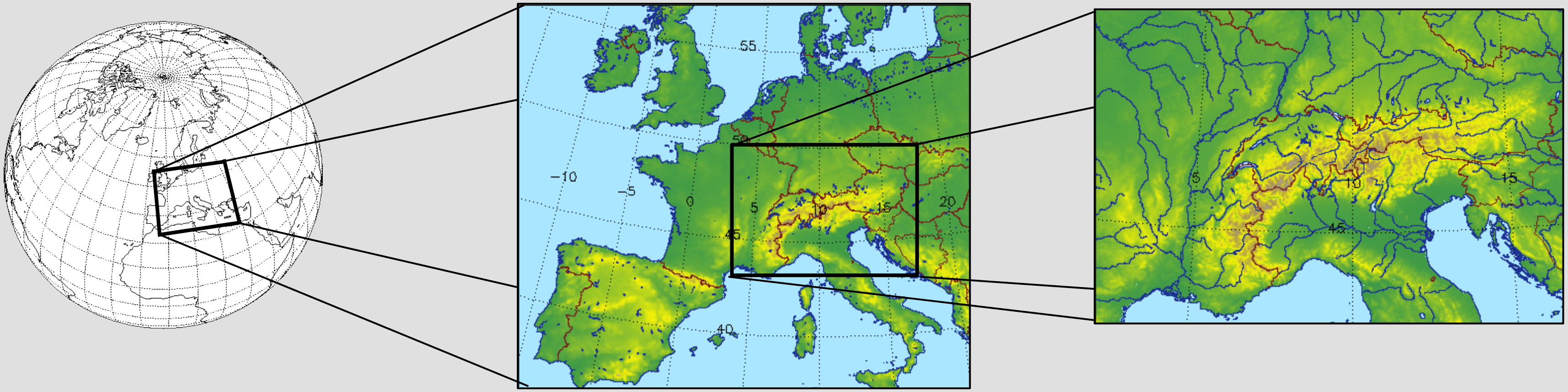
COSMO-2

Δx = 2.2 km, Δt = 20 s

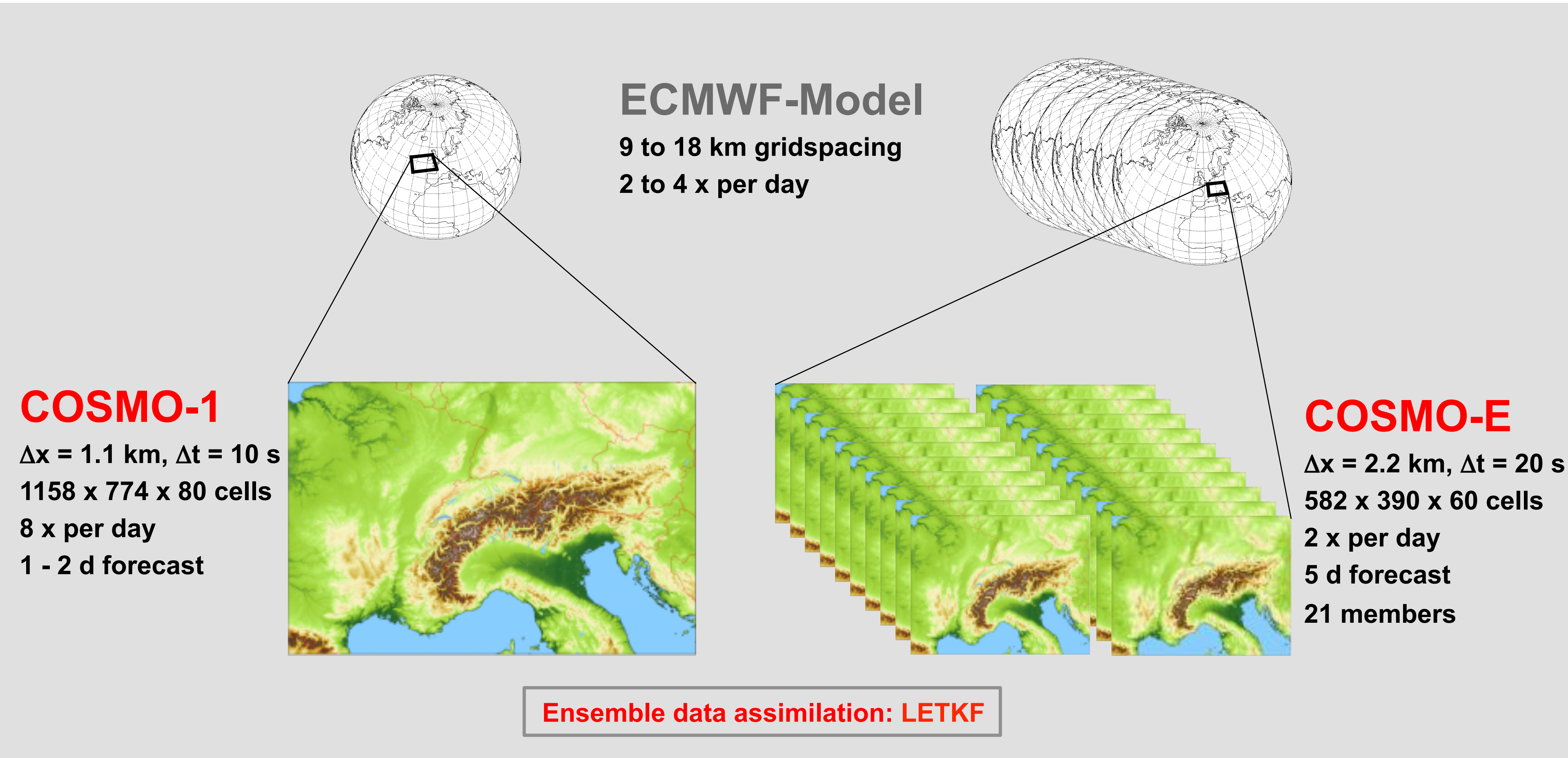
520 x 350 x 60 cells

7 x per day 33 h forecast

1 x per day 45 h forecast



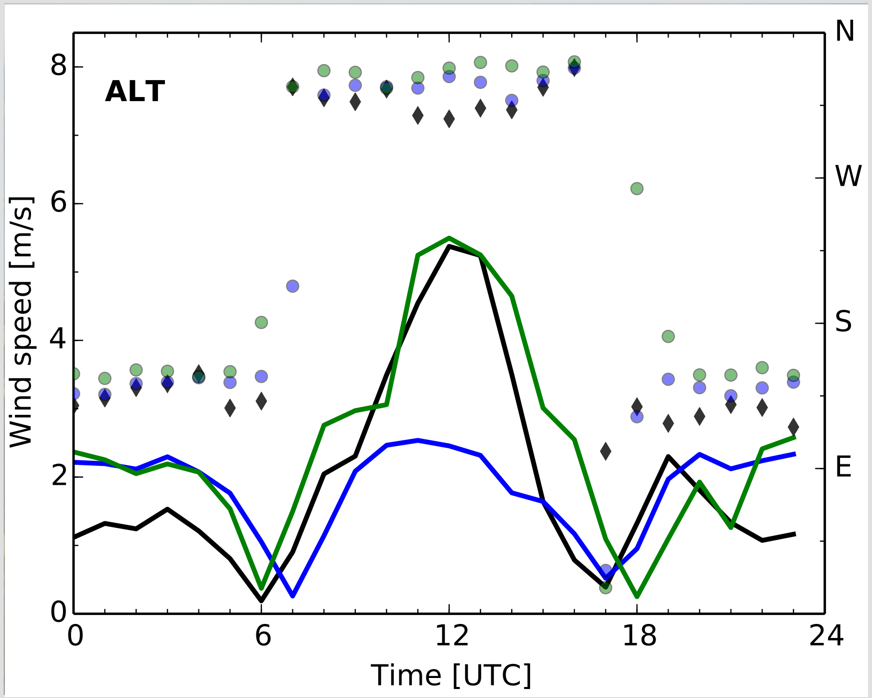
Next-generation system



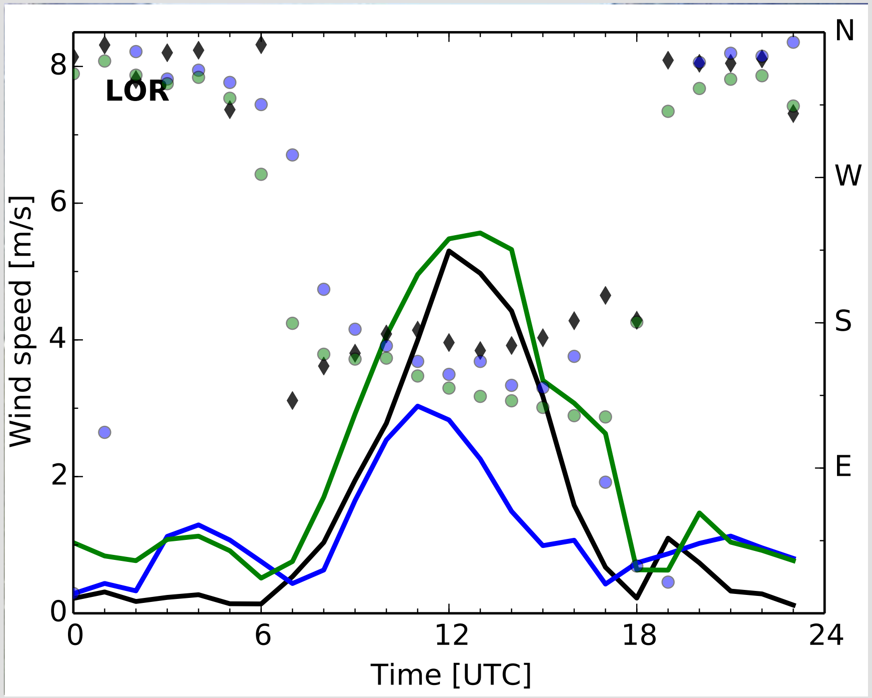
Benefit of high resolution

(18-days for July 9 - 27, 2006)

Altdorf (Reuss valley)



Lodrino (Leventina)

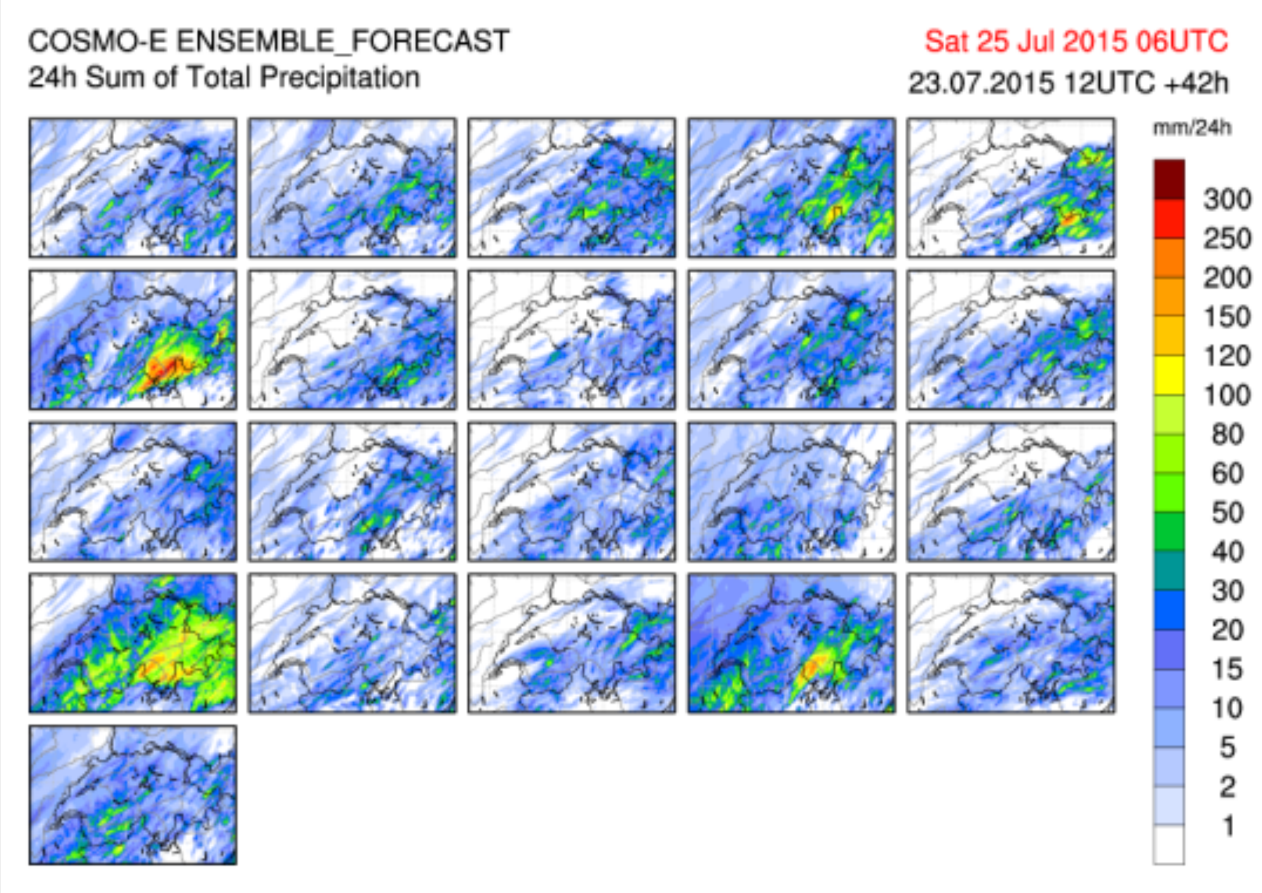


Observation Average wind speed (—) and direction (◊)

COSMO-2 **COSMO-1**

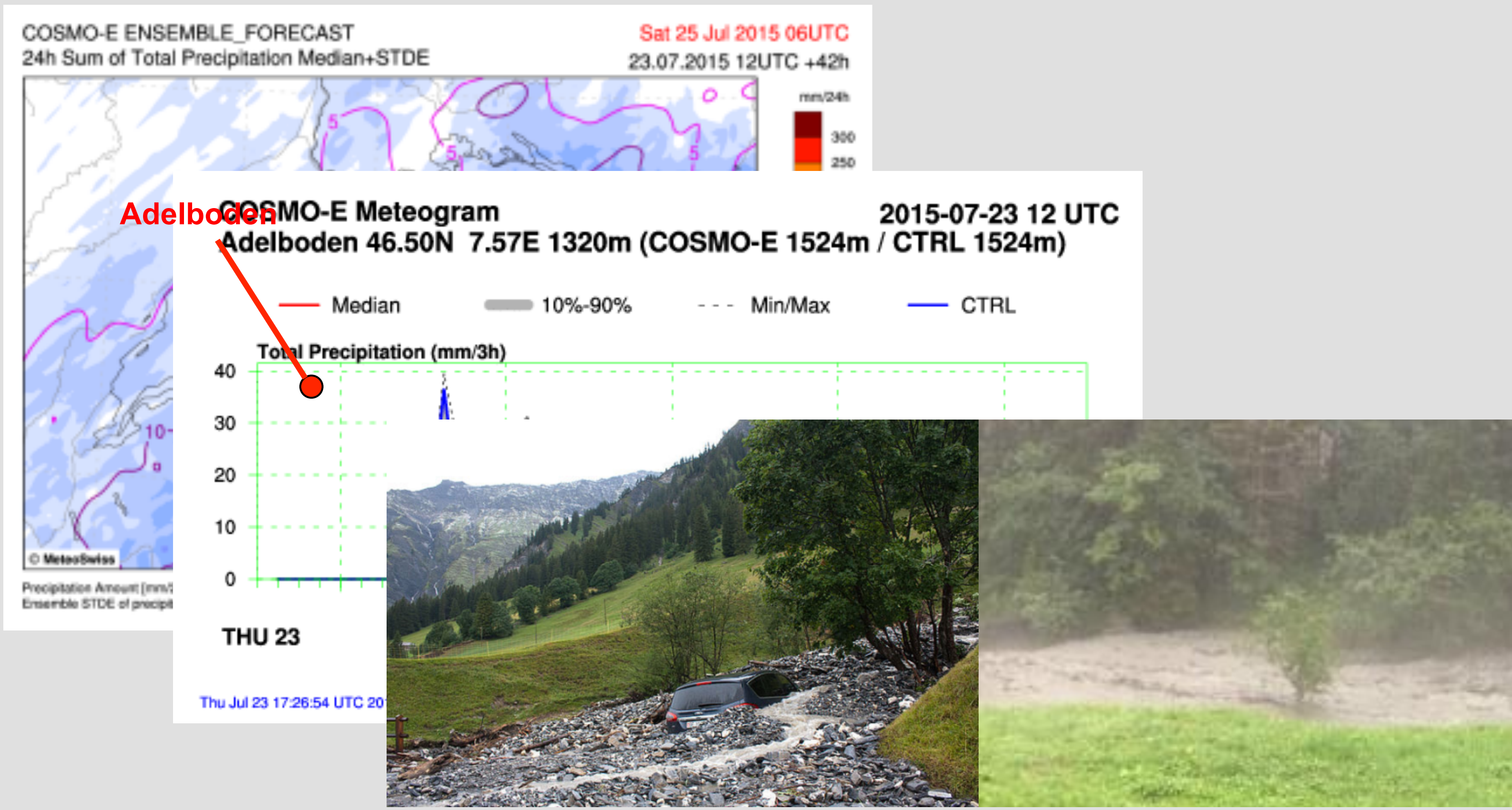
Courtesy of Jürg Schmidli, ETH

Benefit of ensemble
(heavy thunderstorms July 24, 2015)



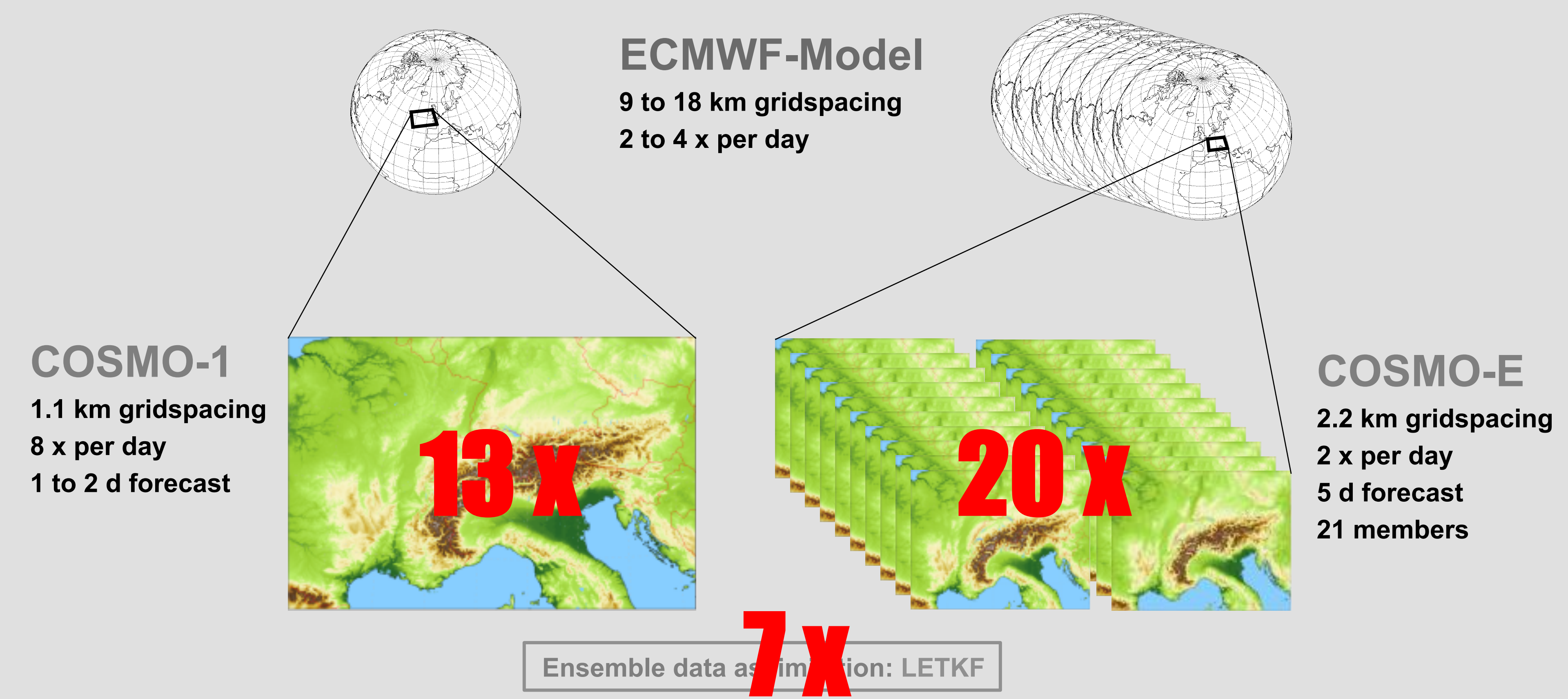
Benefit of ensemble

(heavy thunderstorms July 24, 2015)



Computational cost = 40 x

(relative to current operational system)



Production with COSMO @ CSCS

Cray XE6 (Albis/Lema)

MeteoSwiss operational system

Since ~4 years

Next-generation system

Accounting for Moore's law (factor 4)

Images: CSCS

Co-design: A way out?

Potential

- Time-to-solution driven
- Exclusive usage
- Only one critical application
- Stable configuration
  (code and system)
- Current code is not optimal
- Novel hardware architectures

Challenges

- Community code
  - Large user base
  - Performance portability
  - Knowhow transfer
- Complex workflow
- High reliability
- Rapidly evolving technology
  (hardware and software)

Images: CSCS

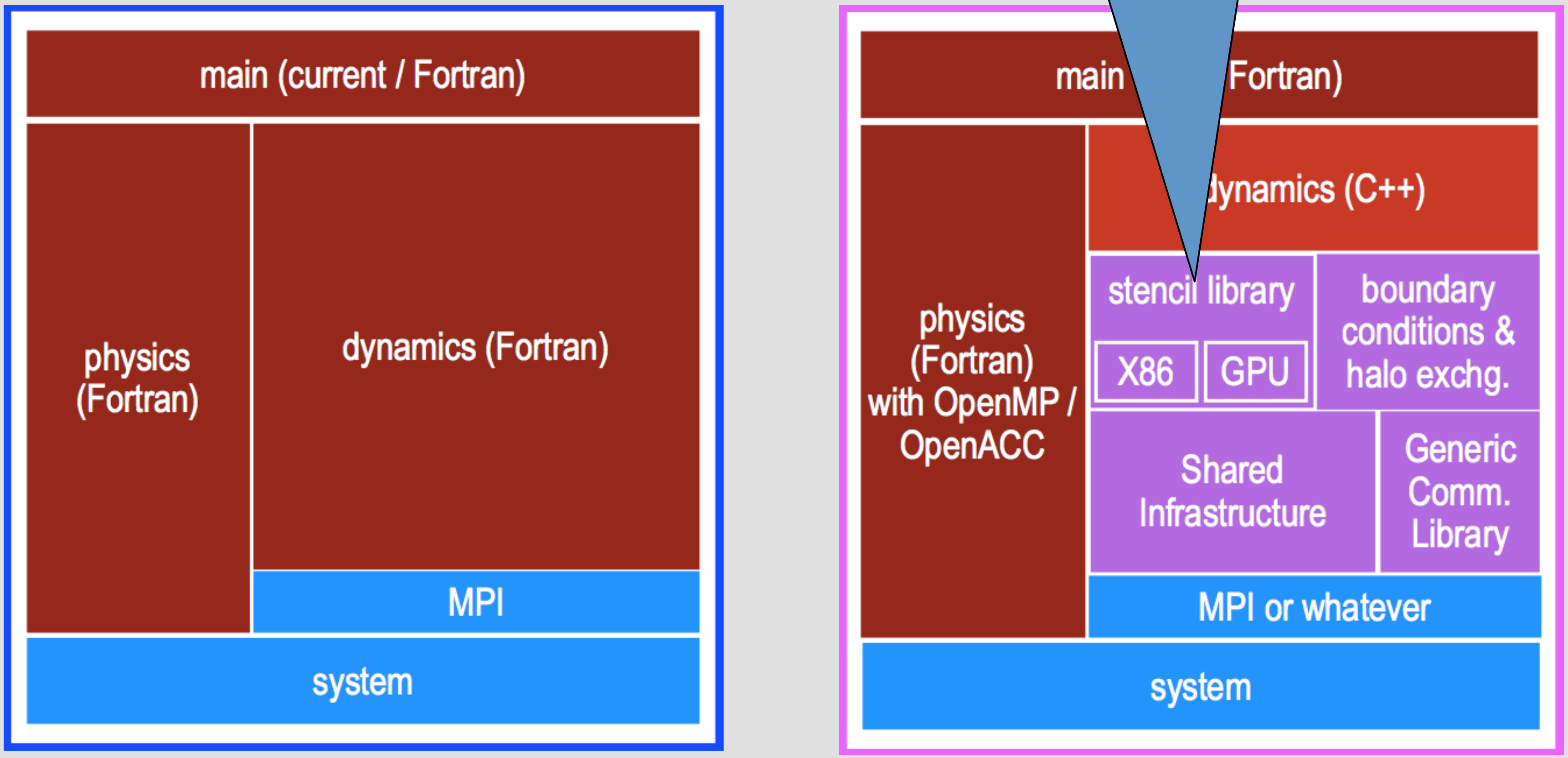
Co-design: Approach

- 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 the HPCN Strategy (HP2C, PASC)

Images: CSCS

Current and new code

We are currently developing a next version of STELLA which is more general (global grids, FEM, …).



adapted from Fuhrer et al. 2014

OpenACC vs. STELLA

- Comparison using hor. diffusion and vert. advection

	runtime	occupancy	DRAM throughput read	DRAM throughput write	shared memory	register usage
non-blocked (naive)						
K20X	0.53 ms	0.266	>75.1 GB/s	>35.5 GB/s	0 B	47-53
K20	0.68 ms	0.285	>39.1 GB/s	>26.3 GB/s	0 B	37-44
blocked						
K20X	0.90 ms	0.283	13.9 GB/s	62.9 GB/s	0 B	73
K20	0.69 ms	0.591	12.7 GB/s	63.1 GB/s	4 B	46
shared						
K20	0.54 ms	0.600	15.9 GB/s	16.1 GB/s	4.272 KB	39
shared-3D						
K20	0.56 ms	0.670	15.4 GB/s	16.1 GB/s	4.272 KB	34
STELLA						
K20X	0.29 ms	0.90				
K20	0.35 ms	0.90				

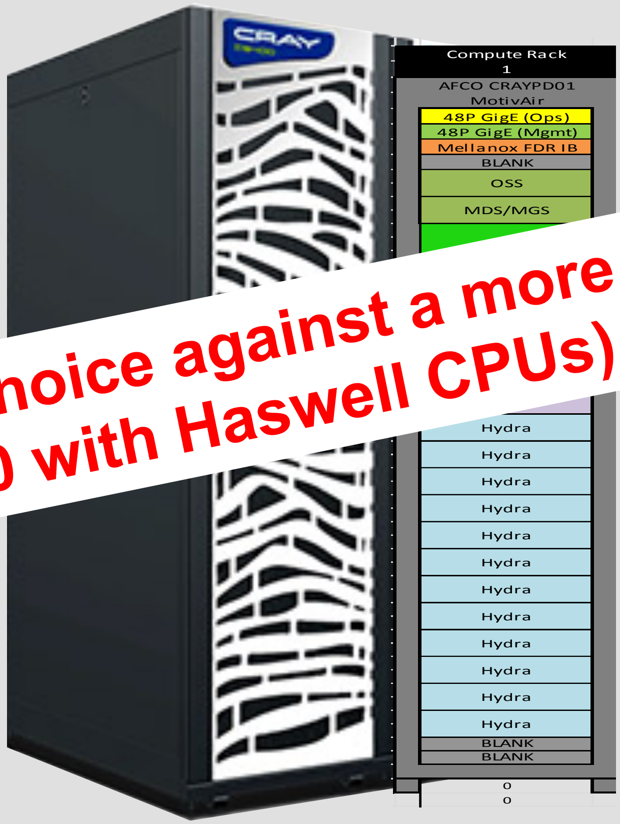
Conclusions

- STELLA implementation is 1.5 – 2.0 x faster
- OpenACC code is portable, but not fully performance portable, many manual optimizations

New MeteoSwiss HPC system

Piz Kesch (Cray CS Storm)

- Installed at CSCS in July 2015
- Public announcement today
- Hybrid system with a mixture of CPUs and GPUs
- “Fat” compute nodes with 2 Int Xeon E5 269 K8
- possible compute odes
- Fully 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)

New MeteoSwiss HPC system

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)

Energy Measurement

- We use power clamp for comparison
- Measurements from PMDB and RUR were within 1% of clamp

Piz Dora (Cray XC40)

- **Power clamp**
  (external measurement which measures wall consumption including AC/DC conversion, interconnect, but excluding blower)
- 1-2 nodes were down and could not be used (considered in computation)
- **PMDB** (1 Hz, per node)
- **RUR** (total per job)

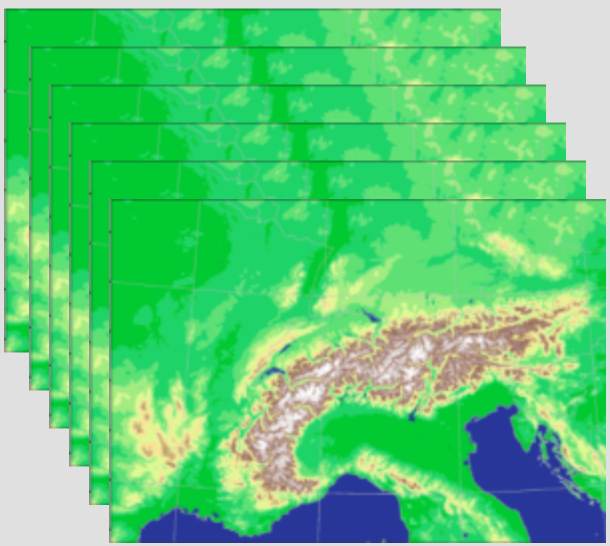
Piz Kesch (Cray CS Storm)

- **Power clamp**
  (external measurement which measures wall consumption including AC/DC conversion, interconnect, but excluding blower)
- Other components (mgmt nodes, extra service nodes, drives) powered down

Benchmark

COSMO-E

2.2 km gridspacing
582 x 390 x 60 gridpoints
120 h forecast



Details

- Planned operational setup by MeteoSwiss
- Required time-to-solution = 2h
  (333 ms per timestep)
- Fill a full rack with members
  (keeping sockets per member constant)
- COSMO v5.0
  (with additions for GPU porting and C++ dynamical core)
- Single precision
  (both CPU and GPU not fully optimized)

Results

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

	Piz Dora	Piz Kesch	Factor
Sockets at required time-to-solution	~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	~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**

„Managment 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 Separating hardware investments from software and workflow investments does not make sense!

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 on a **CS Storm system with fat GPU** nodes starting Q2 2016
- **Energy to solution is a factor 3x smaller** as compared to a “tradiational” 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

/dev/null