DATA ASSIMILATION WITH THE INTEGRATED TERRESTRIAL SYSTEMS PLATFORM TSMP-PDAF

28TH OF FEBRUARY 2020 I HARRIE-JAN HENDRICKS-FRANSSEN[1,2], WOLFGANG KURTZ[4], BIBI NAZ[1,2], SEBASTIAN GEBLER[5], ABOUZAR GHASEMI[3], KLAUS GOERGEN[1,2] AND STEFAN KOLLET[1,2]

[1]Institute of Bio- and Geosciences (IBG-3, Agrosphere), Forschungszentrum Jülich, Jülich, Germany

[2]Centre for High-Performance Scientific Computing in Terrestrial Systems (HPSC-TerrSys), Geoverbund ABC/J, Jülich, Germany

[3]Simulation Laboratory Terrestrial Systems (SimLab TerrSys), Jülich Supercomputing Centre (JSC), Germany

[4]Leibniz Supercomputing Centre, BoltzmannStrasse 1, Garching, Germany

[5]BASF SE, Agricultural Solutions – Global Environmental Fate Modeling, Limburgerhof, Germany

OVERVIEW PRESENTATION

- HPSC-TerrSys and HPC in Earth System Modelling
- Introduction to Terrestrial Systems Modelling Platform (TSMP)
- Introduction to data assimilation
- Data assimilation with TSMP; three examples at different scales
- Conclusions

HPSC TERRSYS

Centre for HPSC in Terrestrial Systems, Geoverbund ABC/J

Simulation Laboratory Terrestrial Systems at JSC

Scientific & Technical Coordination

Appl. Optimization, Parallel Performance

Machine Learning and Data Flows

Model System Maintenance

Participation in research of HPSC TerrSys

Fundamental & Applied HPSC Projects

Earth System Modelling

RCM, LES Simulations

Real-time management in agriculture

Data Assimilation

GRACE

Plant-soil interactions

Hydrogeophysics

PhD Student Organization and Supervision

Training of PhD students

Semi-annual reports

Seminars

International visits & conferences

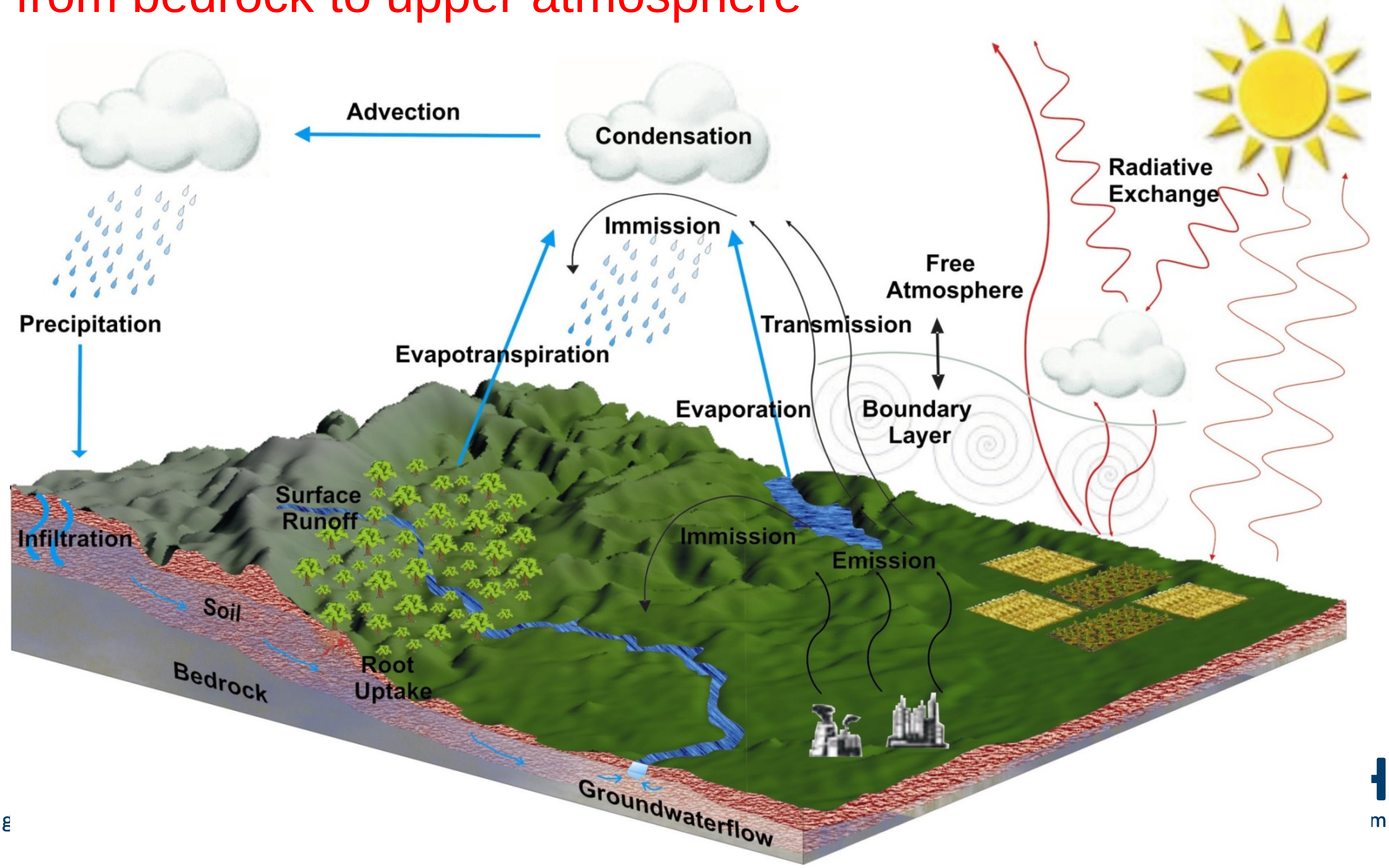
Fall School
Sept. 21.-25., 2020

Scientific Advisory Board (2018-2021)

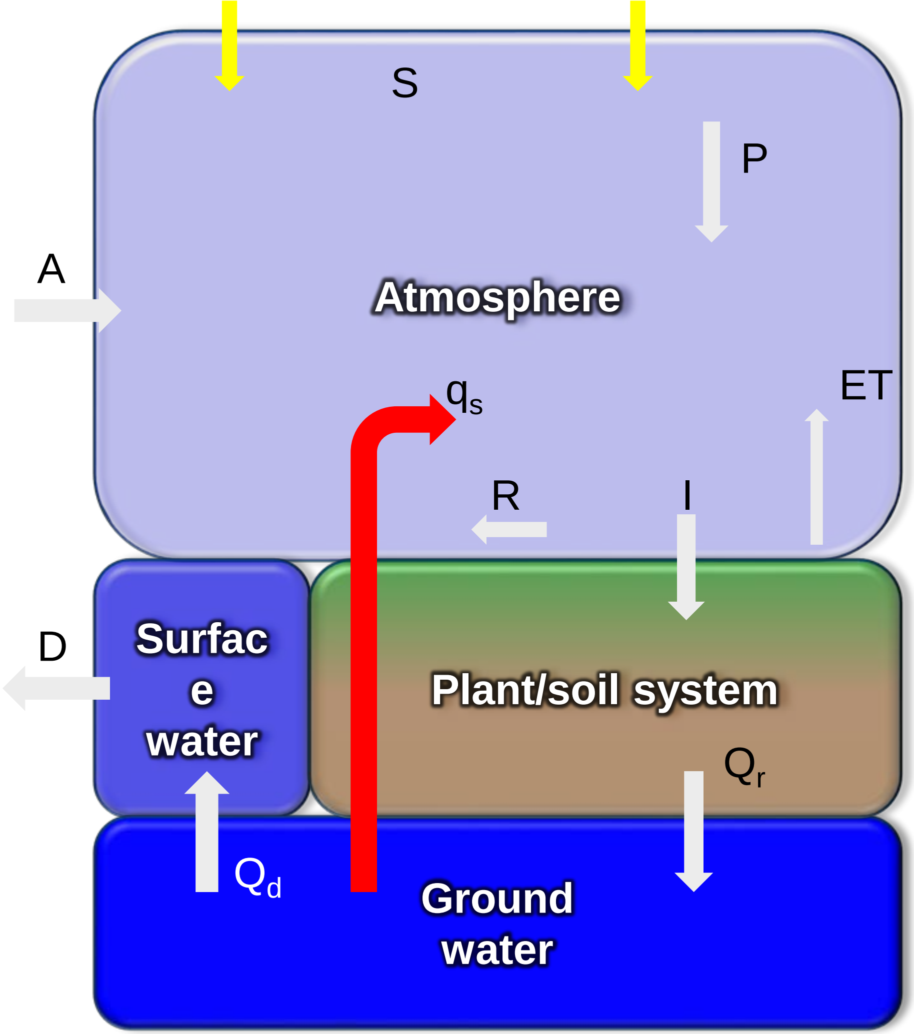
- **Thomas Jung, Alfred Wegener Institute, Bremerhaven – Climate dynamics**
- **Anne Verhoef, University of Reading, United Kingdom – Land surface processes**
- **Michel Kern, Maison de la Simulation (INRIA) – Subsurface modelling**

TERRESTRIAL SYSTEM

Aim: Close water, energy and biogeochemical cycles from bedrock to upper atmosphere



TERRESTRIAL SYSTEM: SOME GOVERNING EQUATIONS



$$\frac{\partial T}{\partial t}+\mathbf{v}\cdot\nabla T=\frac{1}{\rho c_p}\left(\frac{\partial p'}{\partial t}+\dots\right)+\frac{Q}{c_p}$$

$$\frac{\partial q^k}{\partial t}+\mathbf{v}\cdot\nabla q^k=-\frac{1}{\rho}\left(\nabla\cdot\mathbf{J}^k+\dots\right)-\frac{1}{\rho}I^k$$

$$\left(\bar{q}_s-\bar{q}\right)=\frac{E}{k u_* \rho}\left[\ln\left(\frac{z-d_0}{z_0}\right)-\Psi_{sv}(\zeta)\right]$$

$$\frac{\partial \psi_s}{\partial t}=\nabla\vec{v}\,\psi_s-q_r(x)-q_e(x)$$

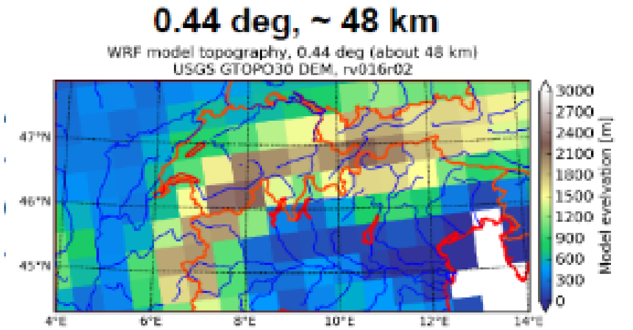
$$S_s S_w\frac{\partial \psi}{\partial t}+\varphi\frac{\partial S_w(\psi)}{\partial t}=\nabla\cdot q+q_s$$

$$q=K_v k_r(\psi)\frac{\partial(\psi+z)}{\partial z}$$

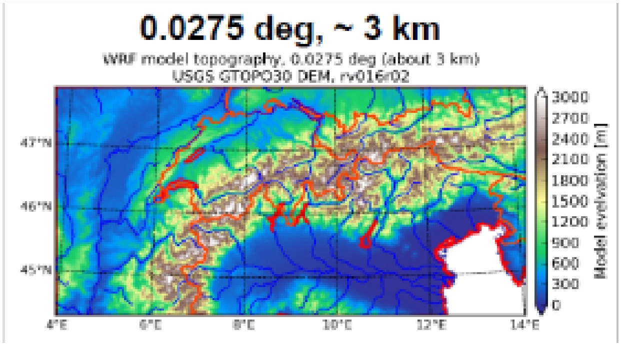
DEVELOPMENTS IN EARTH SYSTEM MODELLING

Convection permitting, “hyper” resolution (added value), short output intervals, big data volumes

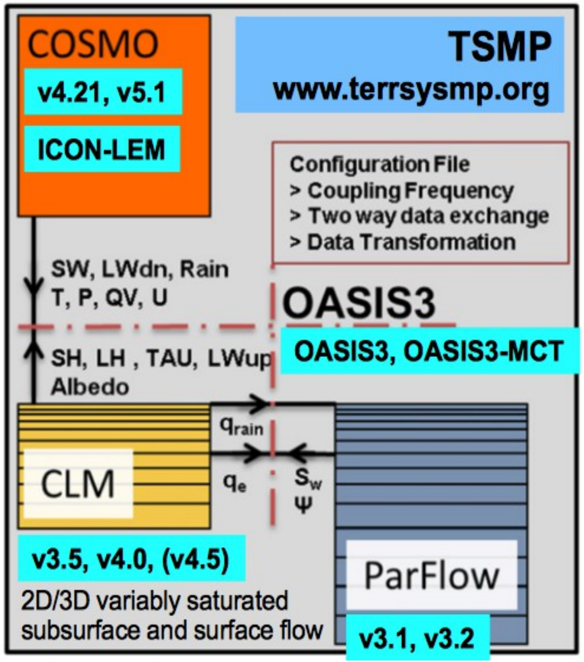
Multiphysics, **fully coupled** (regional) model systems (“Earth system simulator”)



0.44 deg, ~ 48 km



0.0275 deg, ~ 3 km



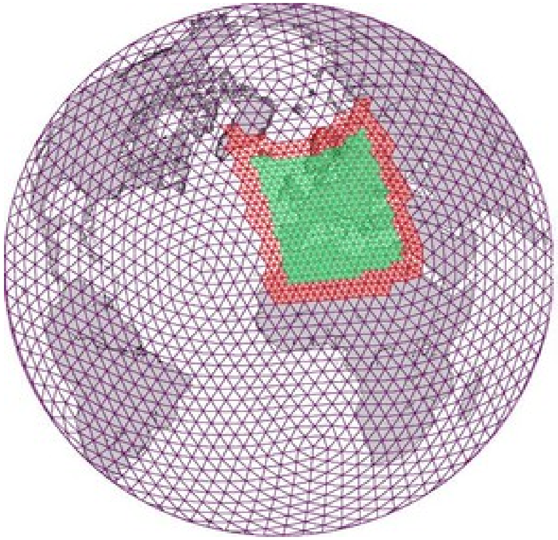
Shrestha et al. (2014, Mon Weather Rev)

- ***Towards* extreme scaling**, global 1km resolution, fully coupled
- Contribution to a more **integrated Earth system science** approach

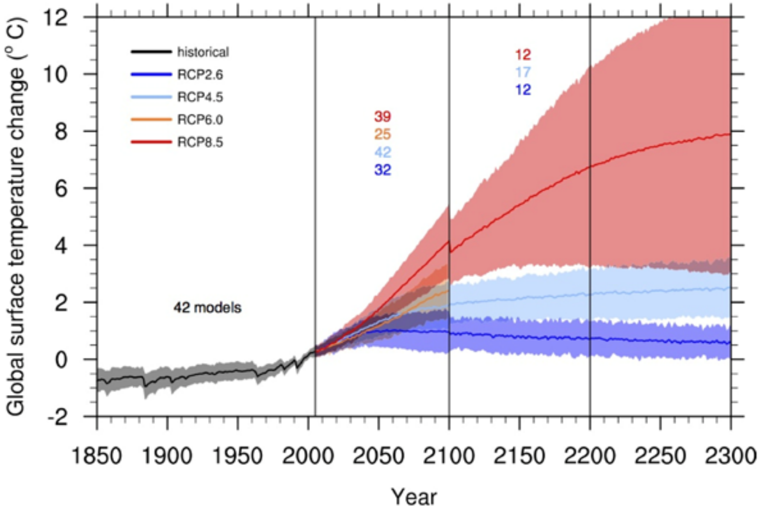
DEVELOPMENTS IN EARTH SYSTEM MODELLING

Increasing domains (multi-scale processes, AMR), **data synthesis**, new data types

Data assimilation (uncertainties), **long integration** times, increasing **ensemble sizes**



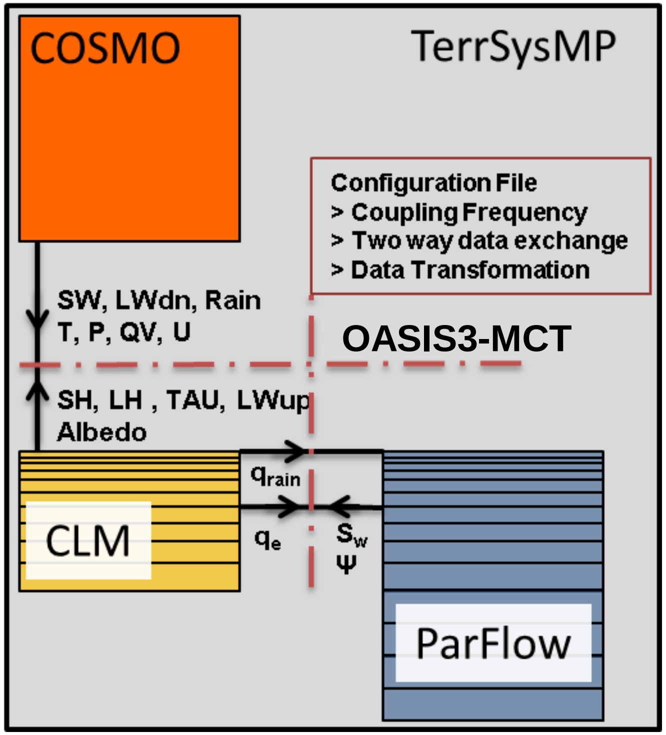
https://www.earthsystemcog.org



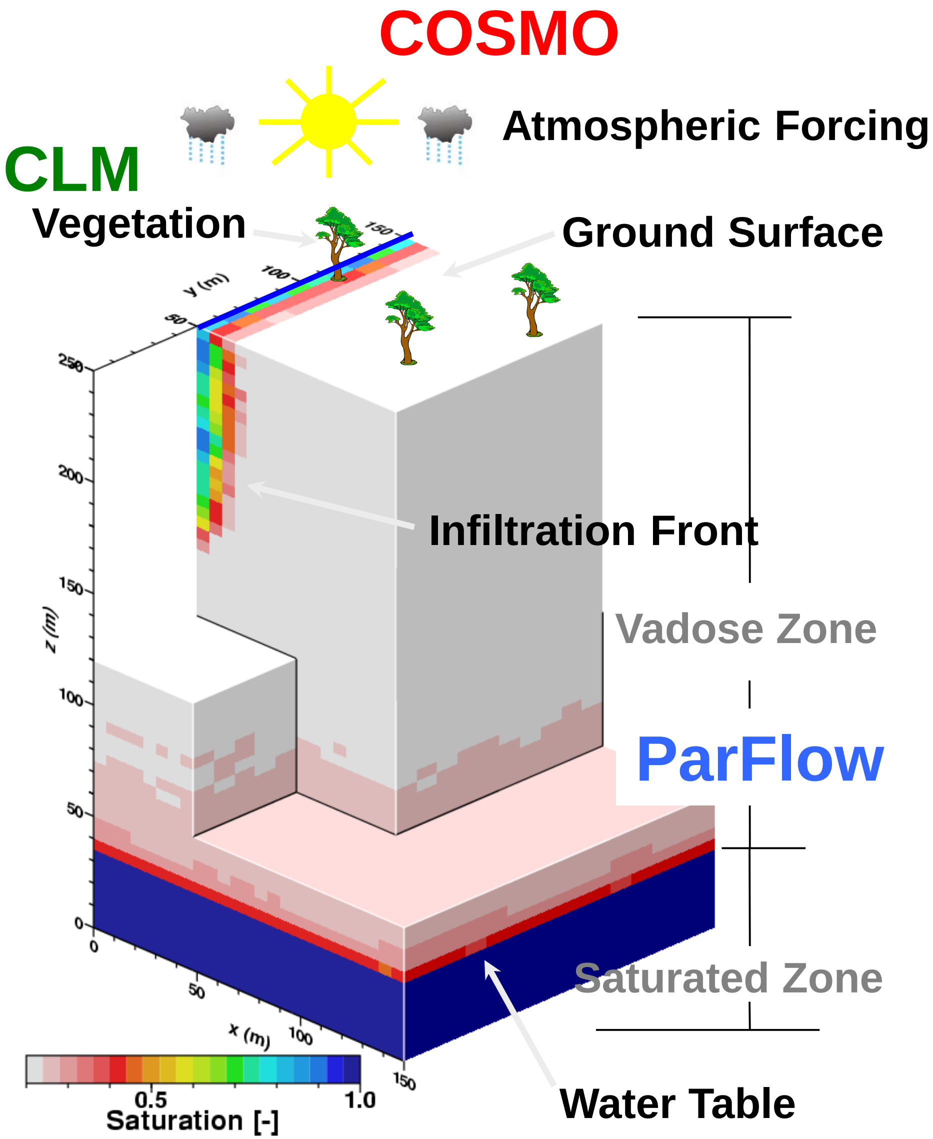
Collins et al. (2013, IPCC WG1 AR5)

- **Hardware** / HPC developments (e.g., GPUs, schedulers); **algorithms** (e.g., solver libraries, memory usage); new **software** / development paradigms (“separation of concerns” via DSLs, in-situ, compression, etc.)

TSMP



TSMP schematic
Shrestha et al. 2014



TSMP

- Highly modular, massively parallel regional Earth system model, extensively profiled.

- Open source code from GitHub (https://github.com/HPSCTerrSys/TSMP) under MIT license, including documentation, pre- and post-processing tools, example test cases.

- It is ported on JURECA and JUWELS, DKRZ-Supercomputer MISTRA and can be ported easily on a single x86 workstation (PC or laptop). Ported using GCC and Intel compilers and MPI implementations.

- Also available through a Linux virtual machine, with ready-to-run TSMP environment for a TSMP-PDAF data assimilation tutorial test case.

WHY ARE EARTH SYSTEM MODEL PREDICTIONS UNCERTAIN?

- **Model structural errors, for example:**
  - Richards equation in land surface models
  - Soil respiration in land surface models: simple black-box concept

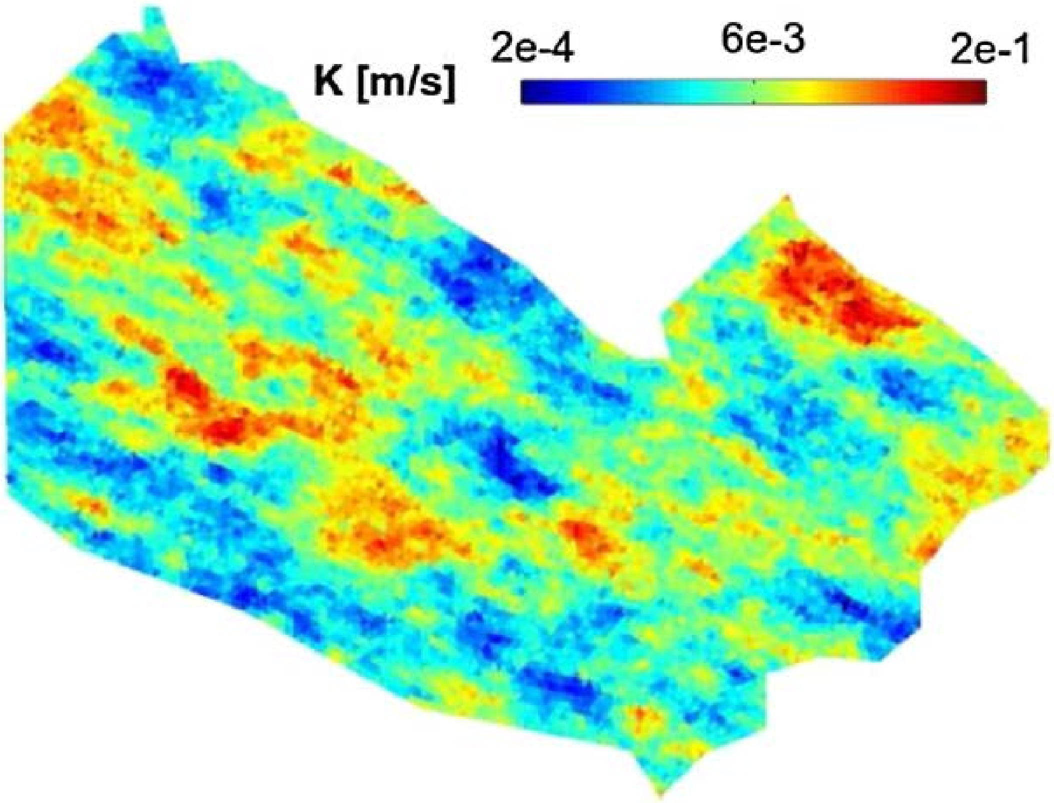
- **Parameter errors, for example:**
  - Soil hydraulic parameters like saturated conductivity
  - Ecosystem parameters like rooting depth

- **Model forcings (for land surface-subsurface), for example:**
  - Precipitation
  - Shortwave radiation

- **Initial conditions, for example:**
  - Initial states of atmosphere like pressure and temperature
  - Soil moisture content

EXAMPLE SUBSUFACE HETEROGENEITY

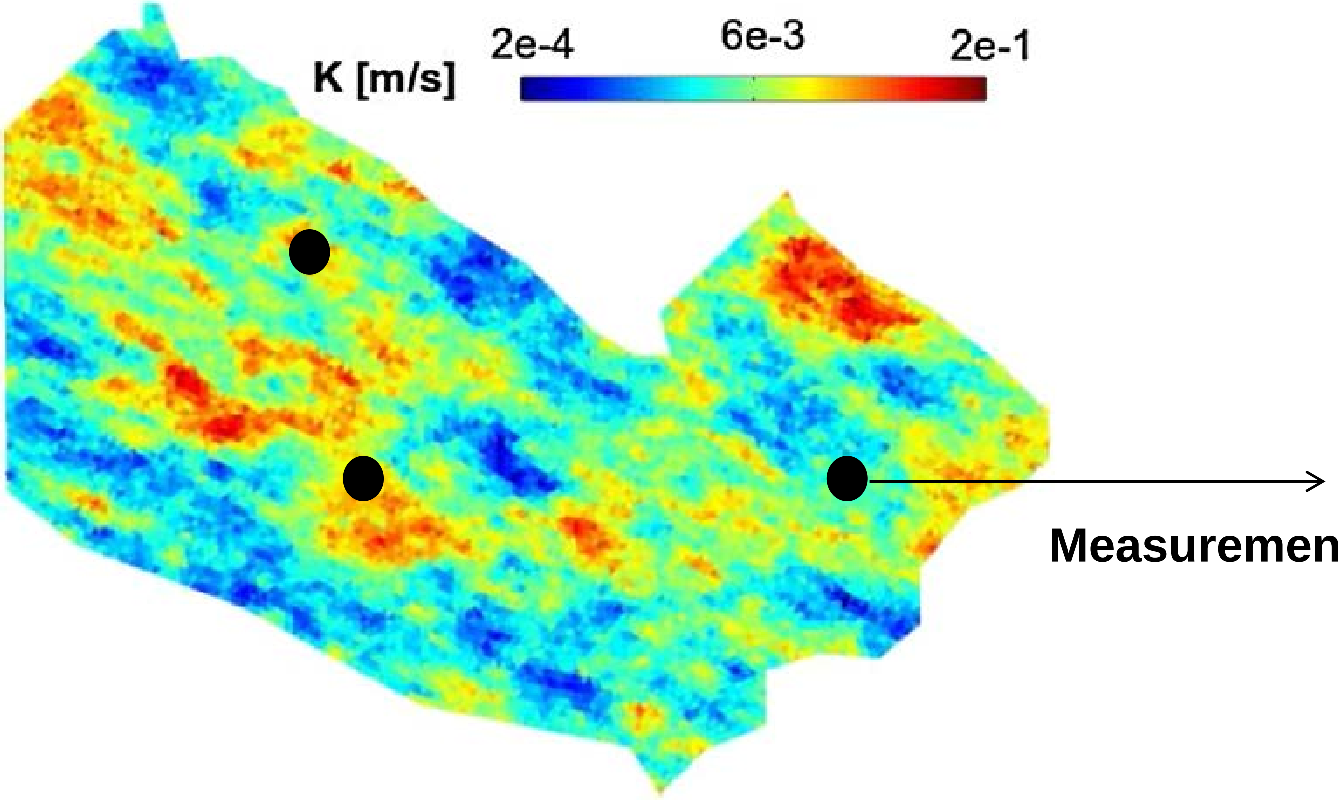
Subsurface heterogeneity could look like this:



JÜLICH
Forschungszentrum

EXAMPLE SUBSUFACE HETEROGENEITY

….. and typically we have limited information:



5 km

ENSEMBLE MODEL CALCULATIONS

- Many sources of (considerable) uncertainty

- Non-linear governing model equations

- Use of mean initial conditions, mean parameter values, mean forcings does not give best estimate of output variables, and of limited value

- Ensemble modelling approach important (not only in geo-sciences)

- Increases requirements for compute resources and parallelization as model runs are X-times repeated (e.g. two parallelization layers)

- Ensemble model runs often at lower spatial resolution

DATA CAN REDUCE UNCERTAINTY

- SYNOP, BUOY, vertical soundings, commercial aircraft, large number of meteorological satellites: ~8 million data per timestep used to correct atmospheric model predictions (data assimilation)

- Much less data available for subsurface: large network of groundwater wells – data spread over institutions
- River discharge data: long time series, but network is reduced
- Networks on soil moisture, land surface fluxes (FLUXNET), ecology (eLTER) established more recently, in last decades
- Increasing number of satellite products available like SMOS and SMAP for soil moisture, MODIS for various variables of interest, GRACE for total water storage, ….

COMBINING MODEL AND DATA

	Sequential approaches (Markov Assumption)	**Batch approaches**
Gaussian approximation	Kalman Filters (KF, EKF, EnKF, and many variants)	Iterative smoothers Variational DA
No Gaussian approximation	Particle Filters	Markov Chain Monte Carlo

ENSEMBLE KALMAN FILTER

$$\mathbf{x}^{t} = M\left(\mathbf{x}^{t-1}, \mathbf{p}, \mathbf{q}\right) + \mathbf{w}^{t}$$

Prediction equation: the model prediction
x = vector with model states
p = vector with parameters
q = vector with model forcings
w = vector with model errors

$$\mathbf{y} = \mathbf{H}\mathbf{x}^{t} + \mathbf{v}^{t}$$

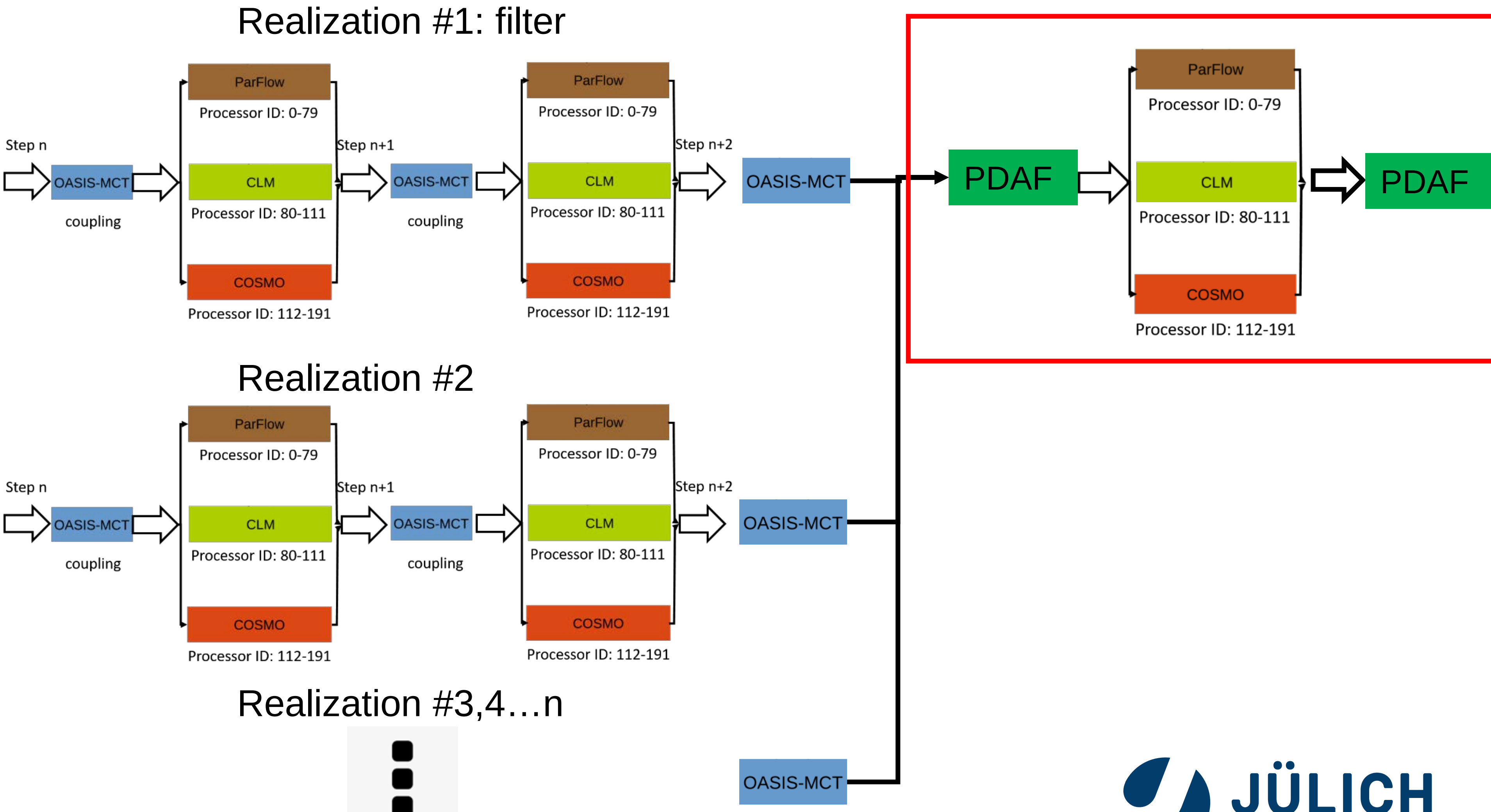
Measurement equation
y = vector with measurement data
H = operator that links measurement and model states
v = vector with measurement errors

$$\mathbf{x}^{t,act} = \mathbf{x}^{t} + \mathbf{K}\left(\mathbf{y} - \mathbf{H}\mathbf{x}^{t}\right)$$

$$\mathbf{K} = \mathbf{C}\mathbf{H}^{\mathrm{T}}(\mathbf{H}\mathbf{C}\mathbf{H}^{\mathrm{T}} + \mathbf{R})^{-1}$$

Analysis equation
K = Kalman gain
C = model covariance matrix
R = measurement error covariance matrix

TSMP-PDAF



Realization #1: filter

Step n
OASIS-MCT
coupling
ParFlow
Processor ID: 0-79
CLM
Processor ID: 80-111
COSMO
Processor ID: 112-191
Step n+1
OASIS-MCT
coupling
ParFlow
Processor ID: 0-79
CLM
Processor ID: 80-111
COSMO
Processor ID: 112-191
Step n+2
OASIS-MCT

PDAF
ParFlow
Processor ID: 0-79
CLM
Processor ID: 80-111
COSMO
Processor ID: 112-191
PDAF

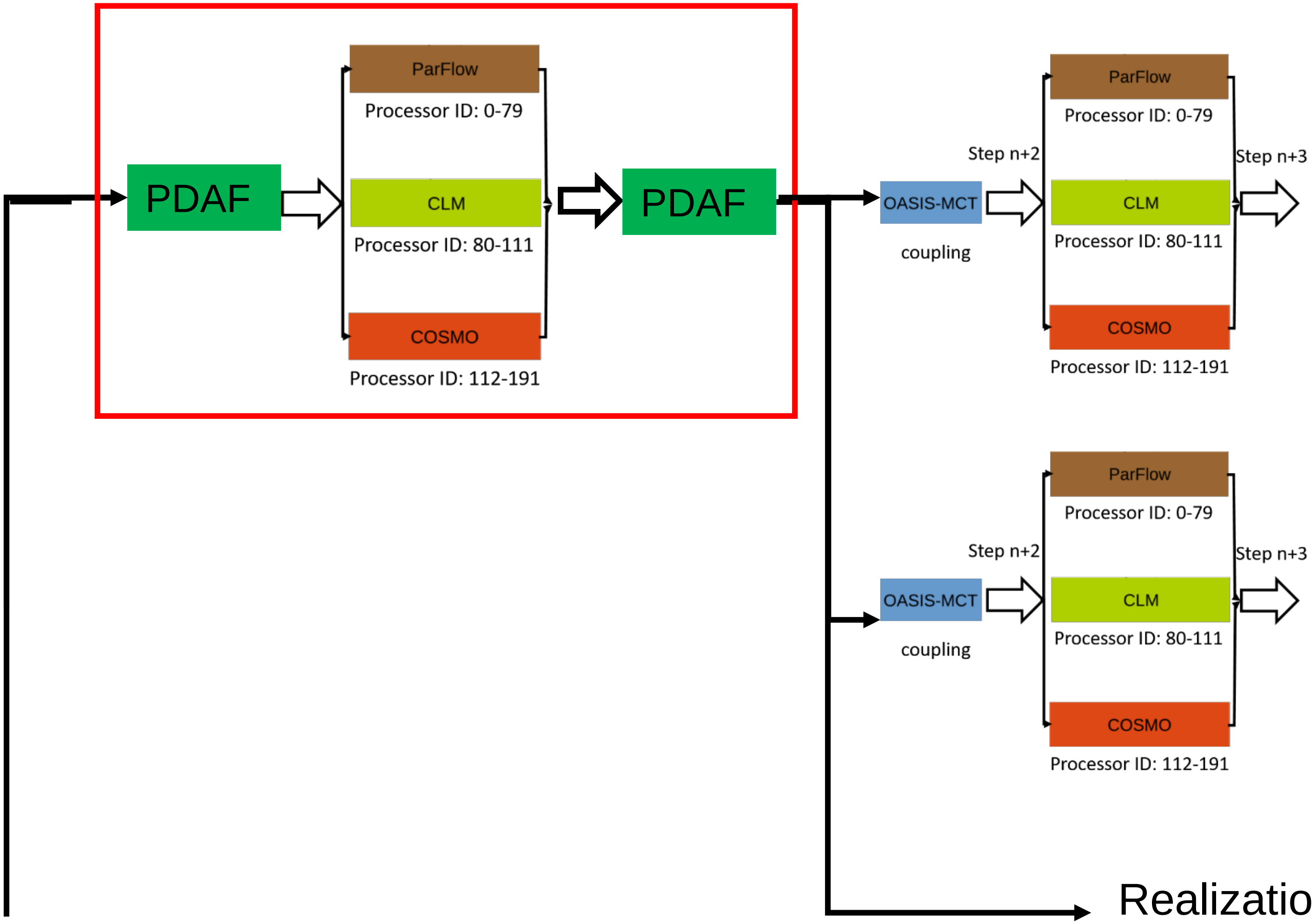
Realization #2

Step n
OASIS-MCT
coupling
ParFlow
Processor ID: 0-79
CLM
Processor ID: 80-111
COSMO
Processor ID: 112-191
Step n+1
OASIS-MCT
coupling
ParFlow
Processor ID: 0-79
CLM
Processor ID: 80-111
COSMO
Processor ID: 112-191
Step n+2
OASIS-MCT

Realization #3,4…n

OASIS-MCT

TSMP-PDAF



PDAF

ParFlow

Processor ID: 0-79

CLM

Processor ID: 80-111

COSMO

Processor ID: 112-191

PDAF

OASIS-MCT

coupling

Step n+2

ParFlow

Processor ID: 0-79

CLM

Processor ID: 80-111

COSMO

Processor ID: 112-191

Step n+3

OASIS-MCT

coupling

Step n+2

ParFlow

Processor ID: 0-79

CLM

Processor ID: 80-111

COSMO

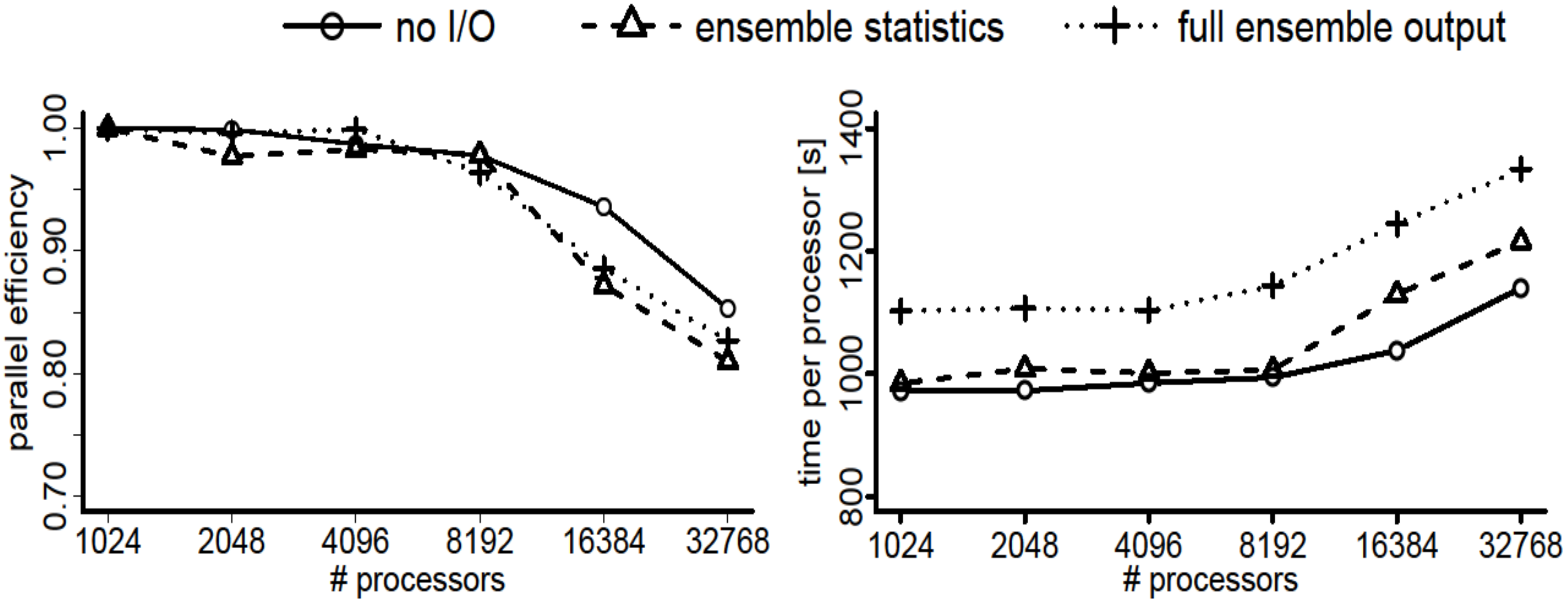
Processor ID: 112-191

Step n+3

Realization #3,4…n

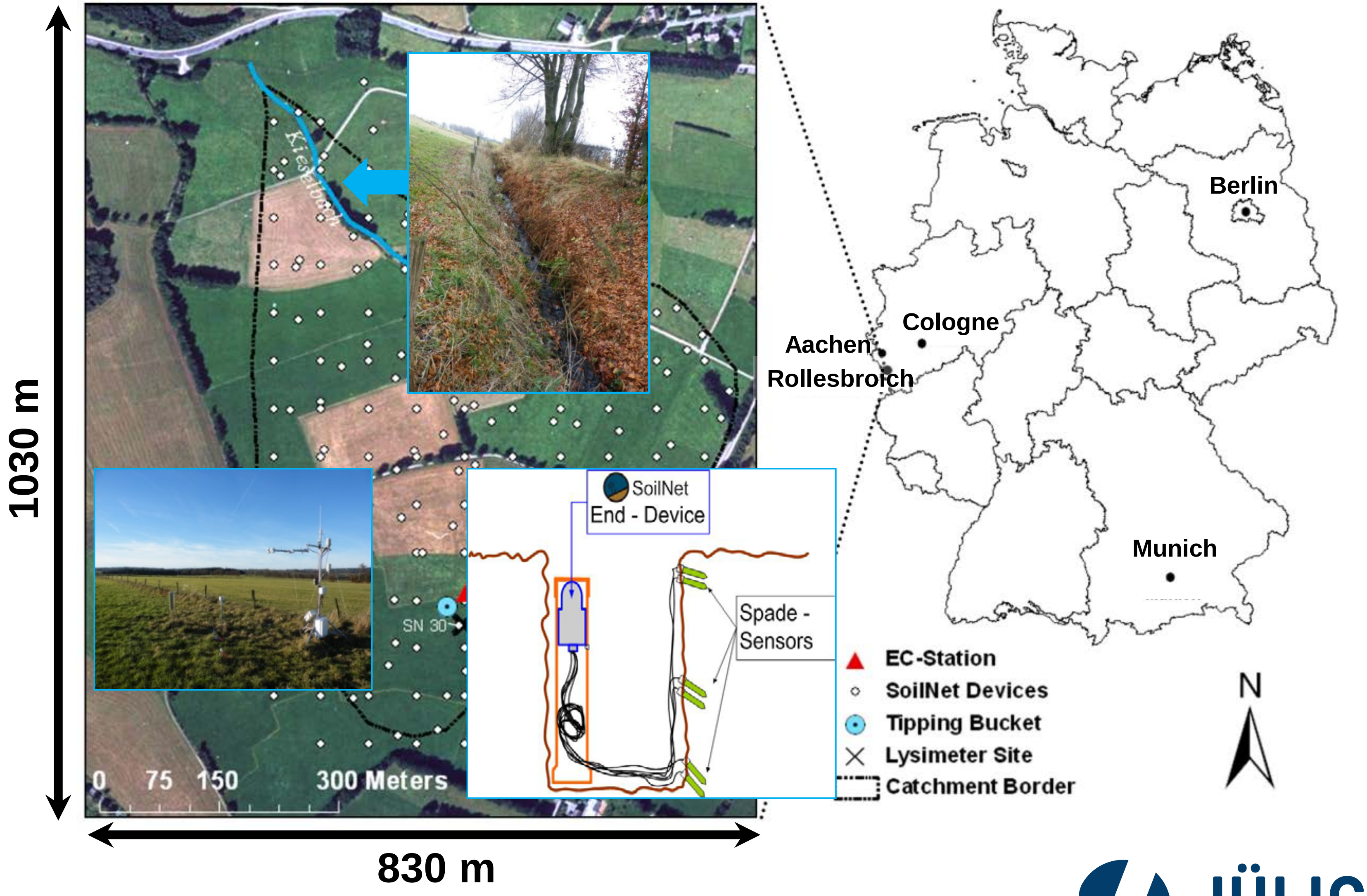
TSMP-PDAF

- PDAF (Nerger and Hiller, 2013) was coupled to TSMP
- COSMO, CLM and ParFlow are parallel, DA in addition also parallel
- DA system is fully integrated (no I/O, no model reinitializations)
- Good scalability through effective use of domain decomposition
- Different DA-algorithms activated (EnKF, local EnKF, LETKF)
- Multiscale SM, GW levels and river water levels can be assimilated



Kurtz et al. (2016), GMD

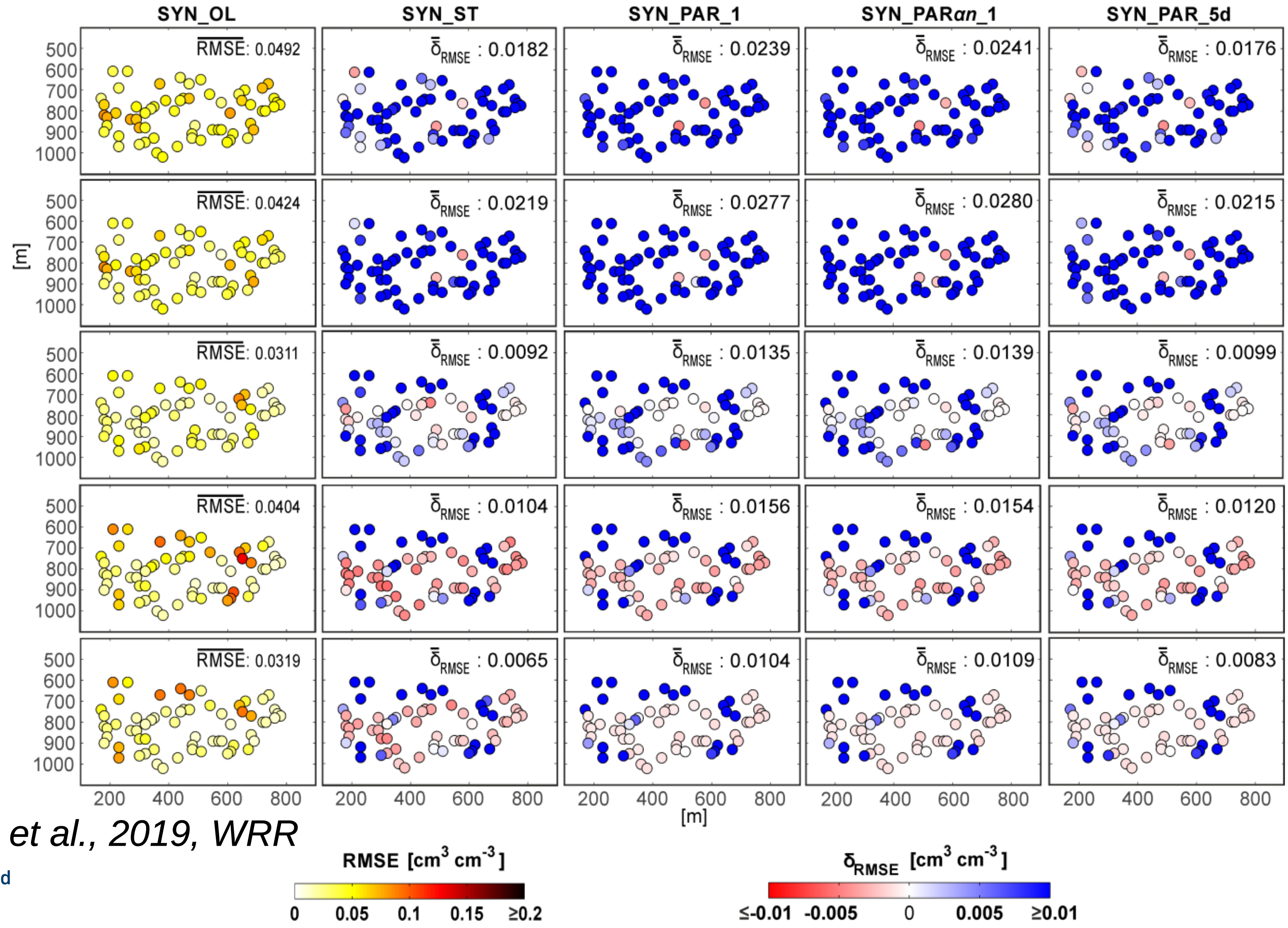
TSMP-PDAF SMALL CATCHMENT (ROLLESBR.)



SET-UP DA STUDY ROLLESBROICH

- Model: CLM-ParFlow-PDAF (from TSMP)
- 128 x 112 grid cells, 10m x 10m resolution
- 20 layers with variable resolution
- Daily soil moisture from 61 sensors, at 5, 20 and 50cm depth assimilated
- 128 ensemble members: precipitation stochastic and 3D heterogeneous fields of soil hydraulic parameters
- Simulation period: May 2011- December 2011
- Real-world experiments and synthetic experiments which mimic real-world

SOIL WATER CONTENT (VERIFICATION)


-33 %

-46%

-46%

-35%

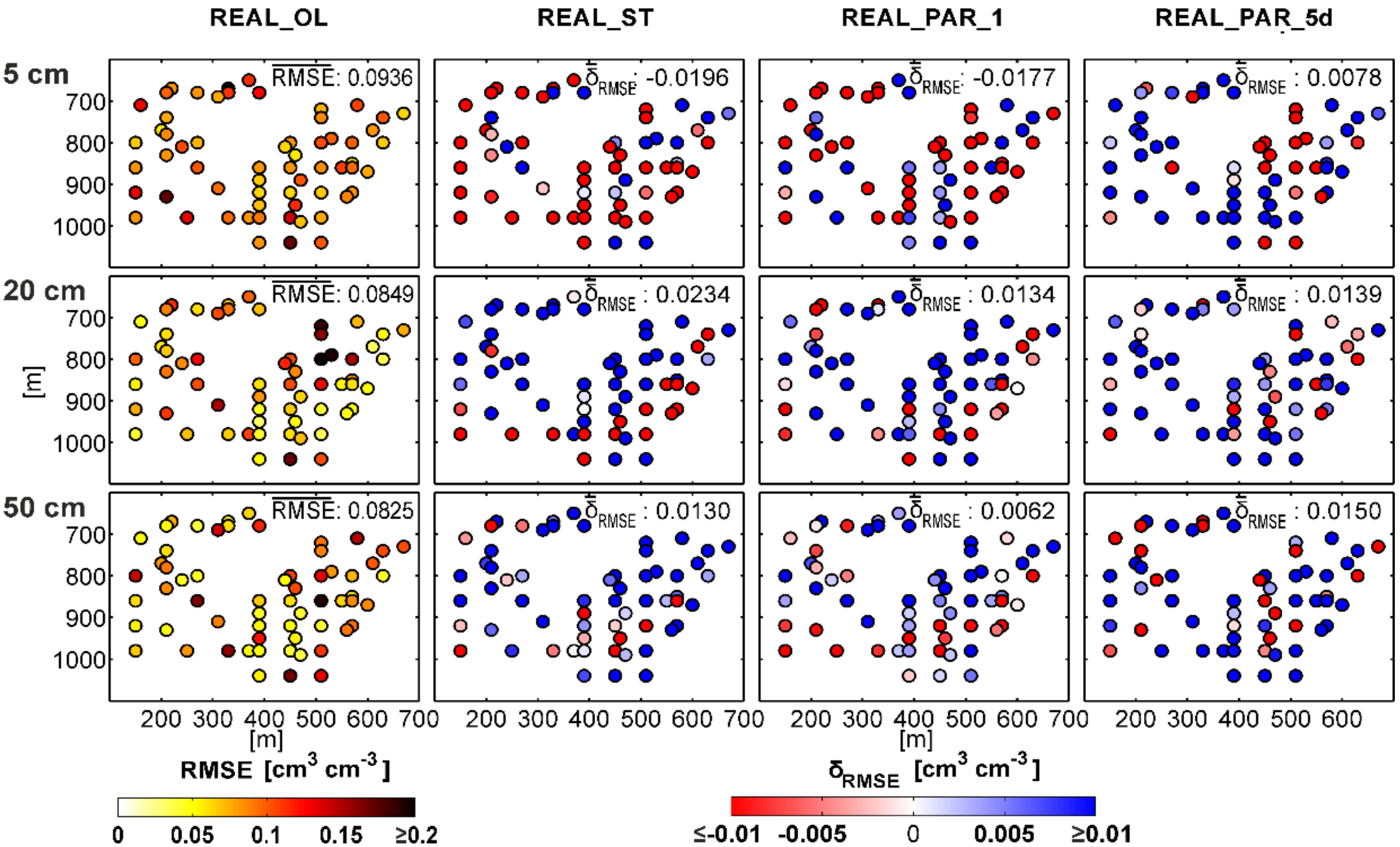
Gebler et al., 2019, WRR

SOIL WATER CONTENT – REAL-WORLD

-7 %

-1%

-14%



Mitglied der Helmholtz-Gemeinschaft

JÜLICH
Forschungszentrum

DISCHARGE ROLLESBROICH

- Only soil moisture was assimilated, not discharge!

- Synthetic case:
  - NSE: -0.03 (open loop) and +0.61 (DA with parameter estimation)
  - Bias: +78% (open loop) to -6% (DA with parameter estimation)
  - RMSE: 14,0 m^3/h (open loop) to 7,3 m^3/h (DA with par. est.)

- Real-world case:
  - NSE: +0.49 (open loop) and +0.67 (DA with parameter estimation)
  - Bias: +65% (open loop) to -24% (DA with parameter estimation)
  - RMSE: 15,8 m^3/h (open loop) to 9,2 m^3/h (DA with par. est.)

- Again clearly better results for synthetic case

TSMP-PDAF LARGE CATCHMENT SCALE

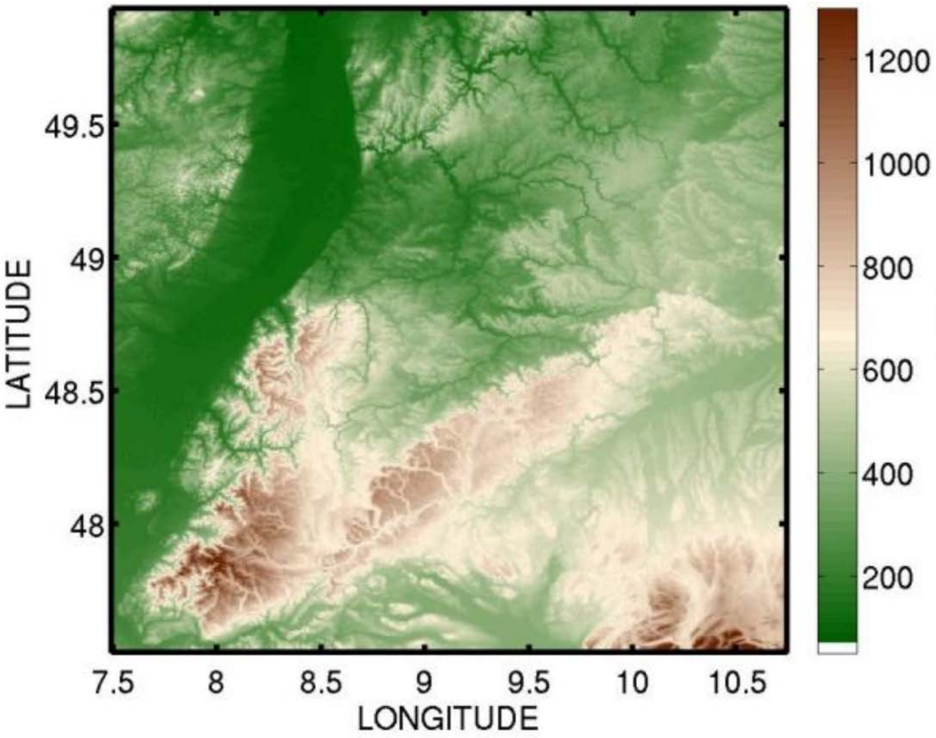
- Virtual reality created with TSMP: COSMO-CLM3.5-Parflow
- Mimics Neckar catchment: 400m resolution, 50 soil layers, 2007-2015.

- Data assimilation experiments with this VR:
  - 800m resolution CLM3.5-Parflow models, year 2015
  - 64 atmospheric forcing ensemble members of four correlated variables (precip, T2M, incoming SW, incoming LW) with space-time geostatistics. Each variable different correlations in space and time.
  - 64 ensemble members for LAI and soil properties

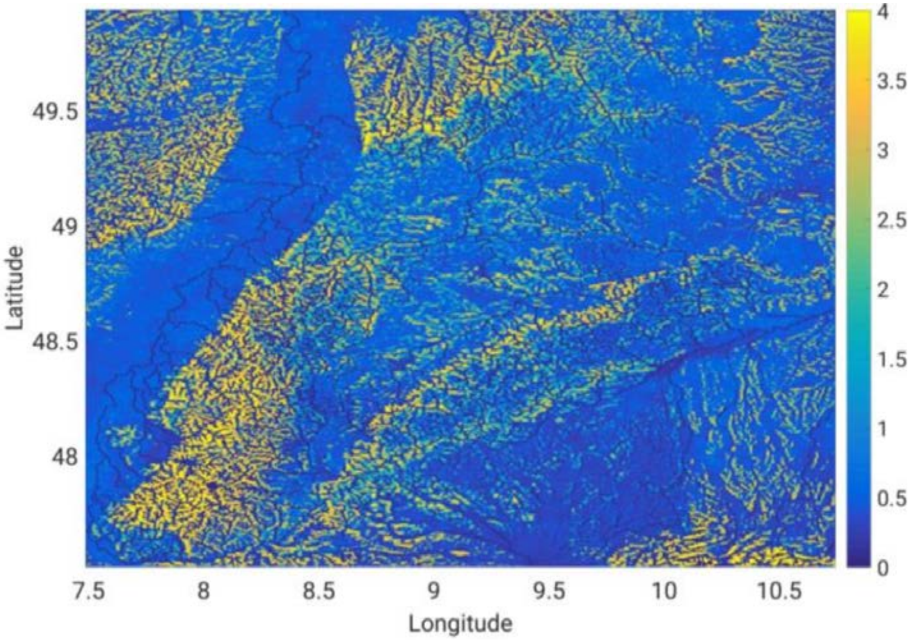
- Soil moisture data (at 5 or 50cm depth) assimilated with EnKF, with/without localization and with/without parameter estimation

Work by Ching-Pui Hung (IBG-3), Bernd Schalge (U Bonn)

VIRTUAL REALITY (VR) NECKAR CATCHMENT

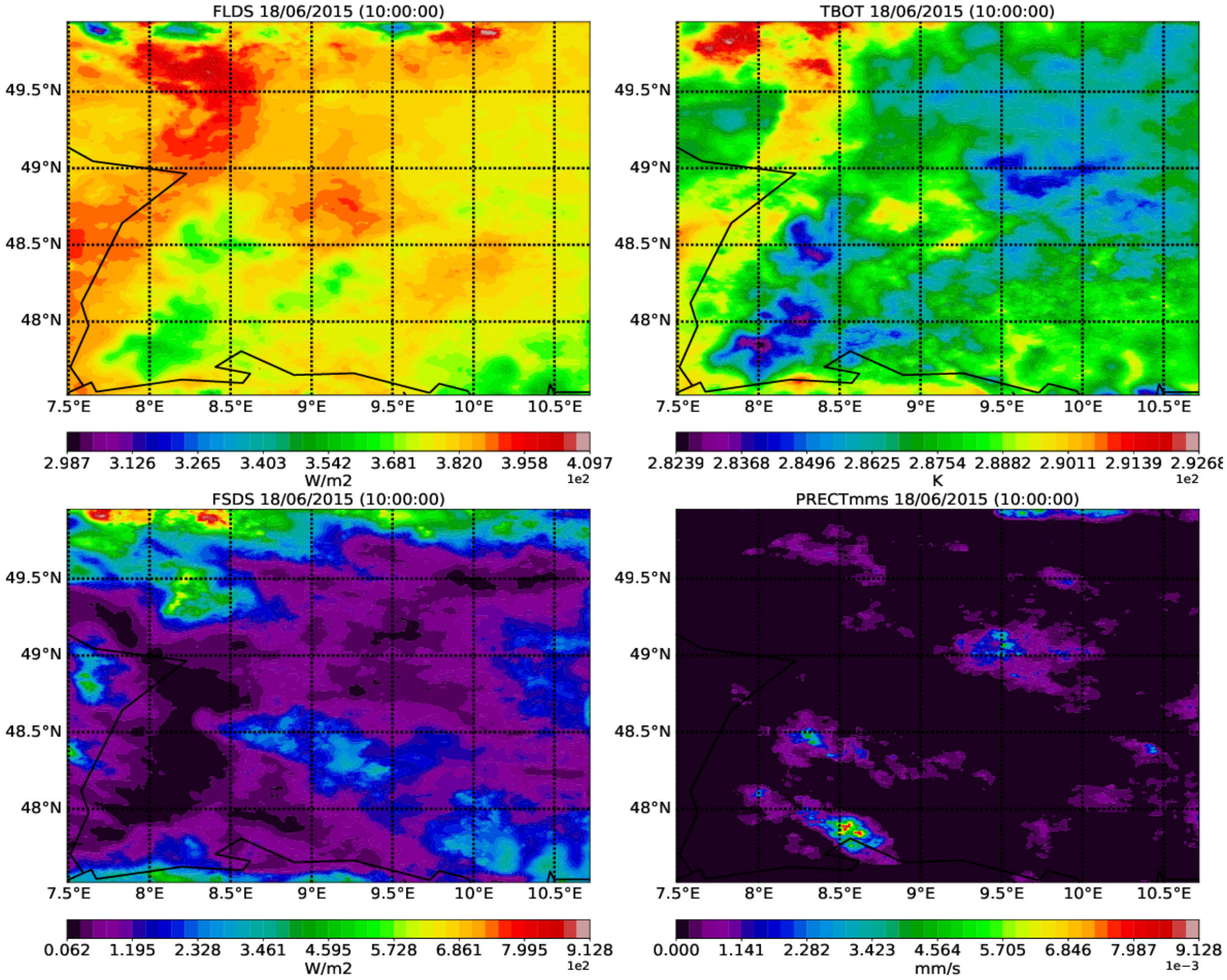


Altitude



Water table depth

VR: ATMOSPHERIC FORCINGS



FLDS 18/06/2015 (10:00:00)

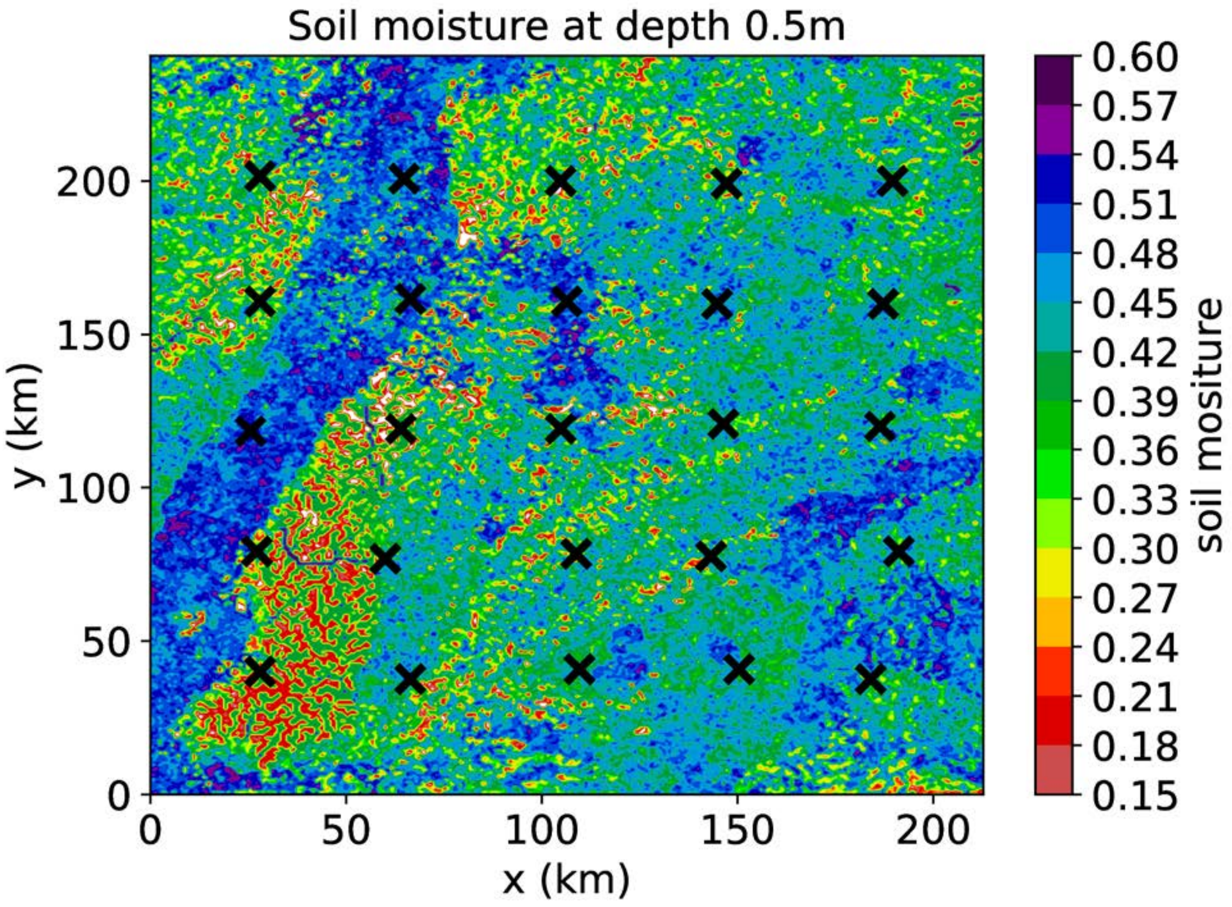
TBOT 18/06/2015 (10:00:00)

FSDS 18/06/2015 (10:00:00)

PRECTmms 18/06/2015 (10:00:00)

Snapshot atmospheric forcings used in land surface-subsurface simulations.

SOIL MOISTURE OBSERVATIONS VR

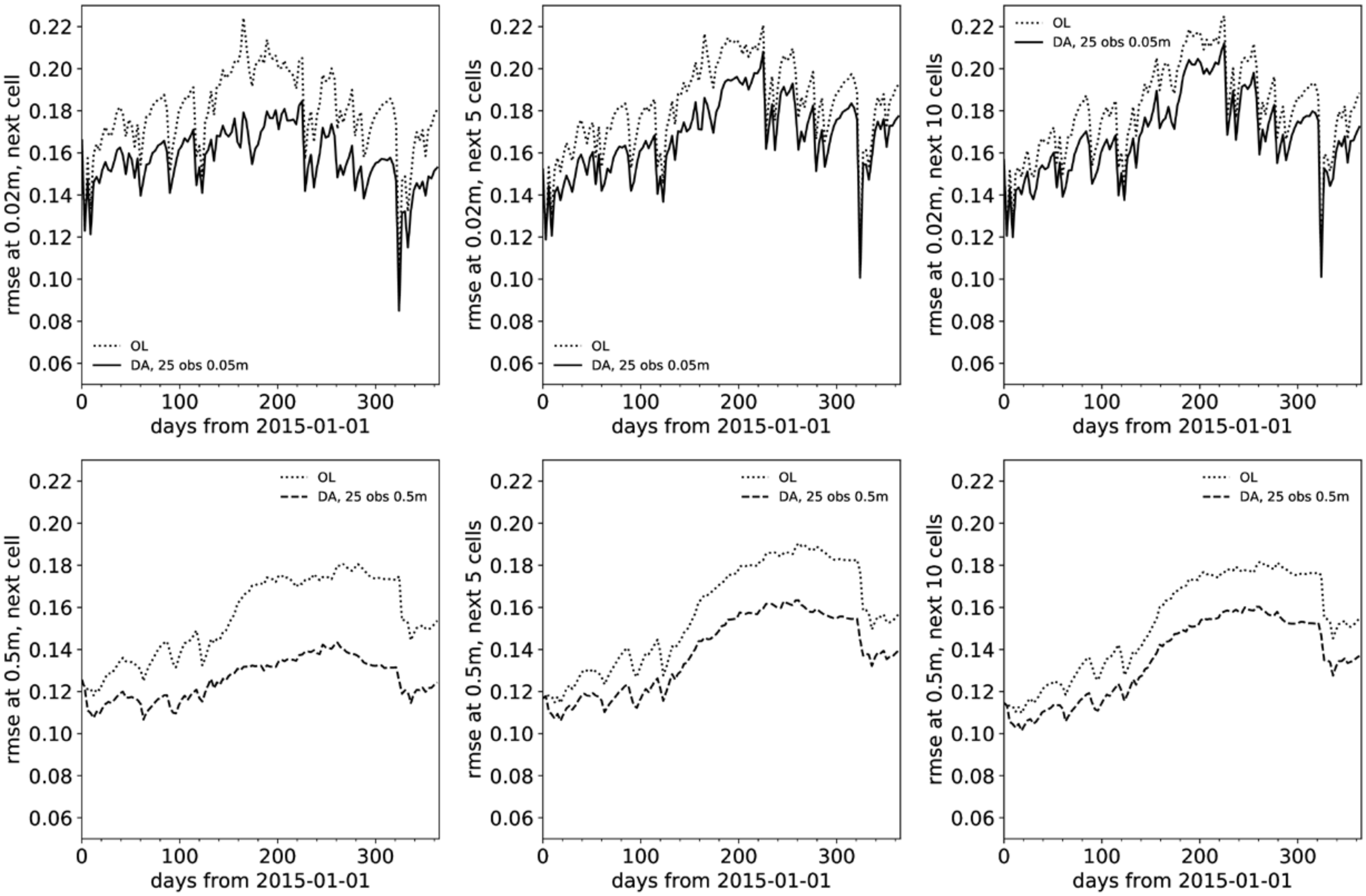


RMSE SOIL MOISTURE CLM-PARFLOW

800m around observ. 4km around obsev. 8km around observ.

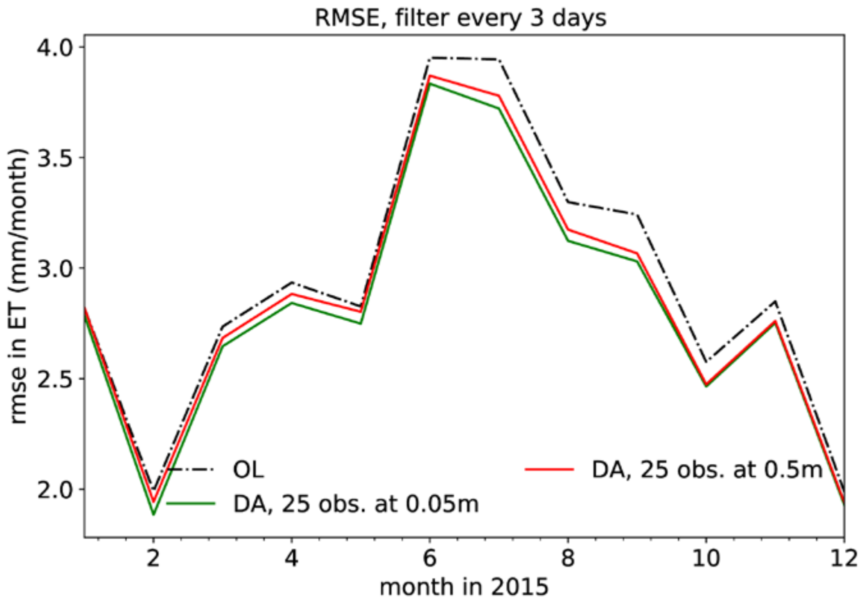
2cm depth

50cm depth

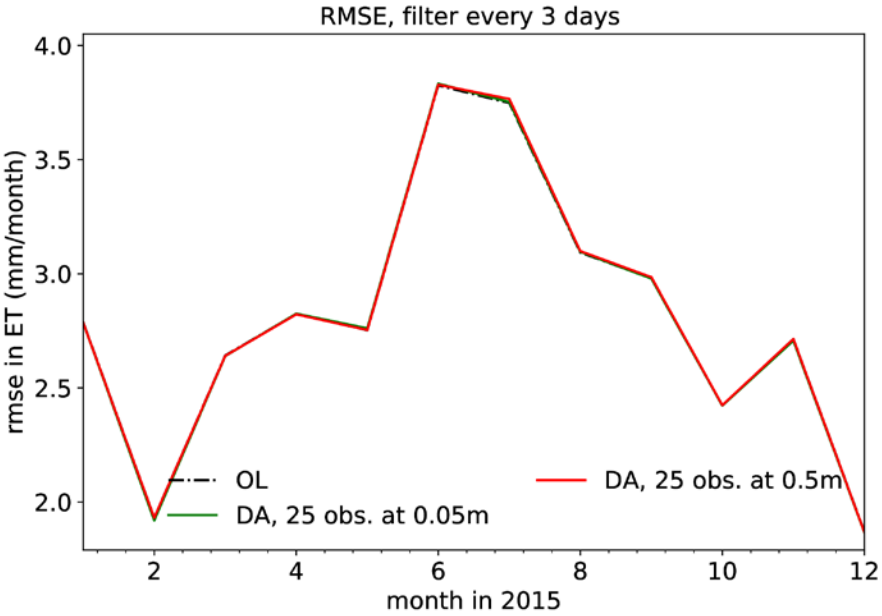


RMSE-REDUCTION BY DA (CLM-PARFLOW)

RMSE EVAPOTRANSPIRATION

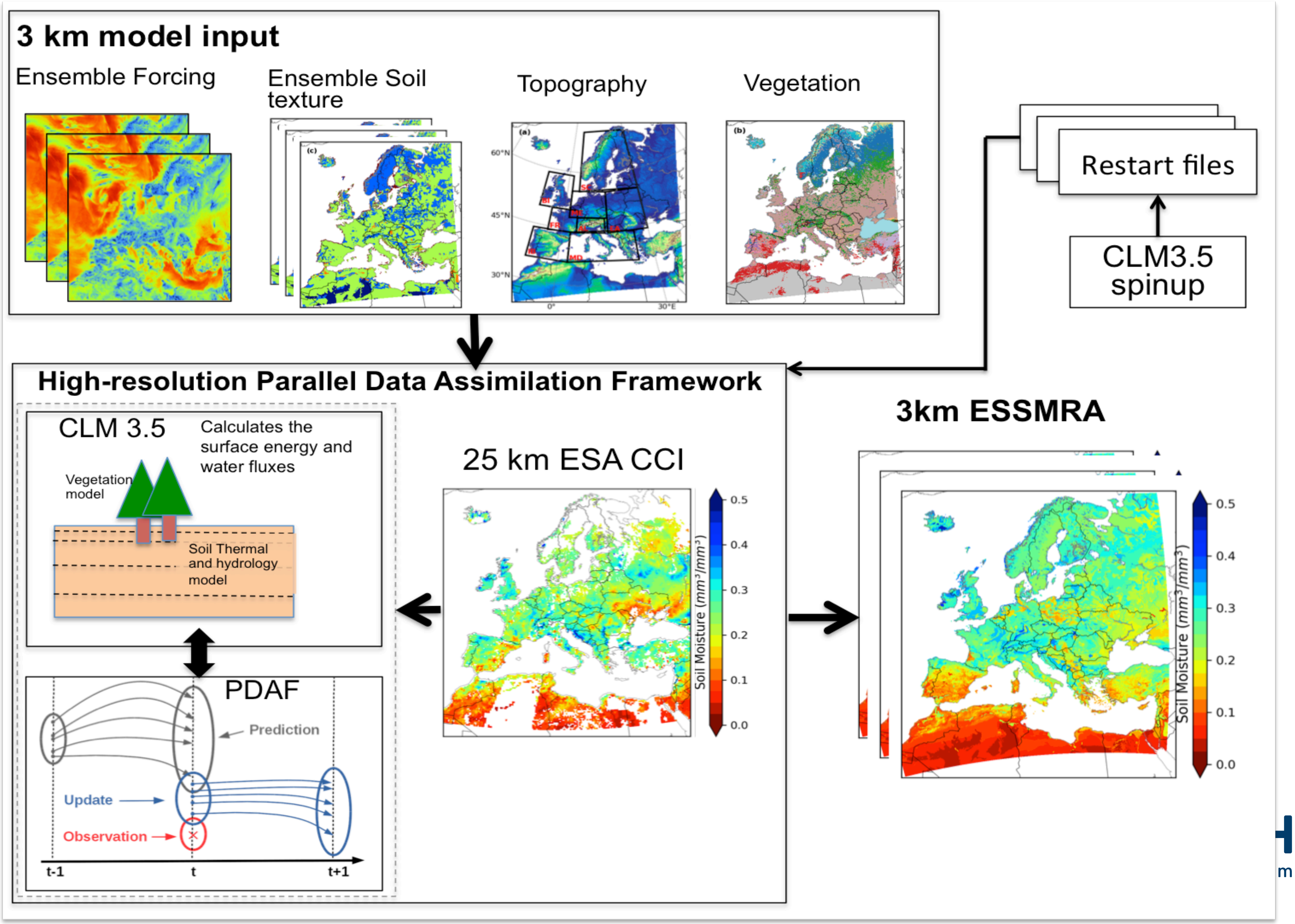


CLM-ParFlow



CLM

TSMP-PDAF CONTINENTAL SCALE



3 km model input

Ensemble Forcing

Ensemble Soil texture

Topography

Vegetation

Restart files

CLM3.5 spinup

High-resolution Parallel Data Assimilation Framework

CLM 3.5

Calculates the surface energy and water fluxes

Vegetation model

Soil Thermal and hydrology model

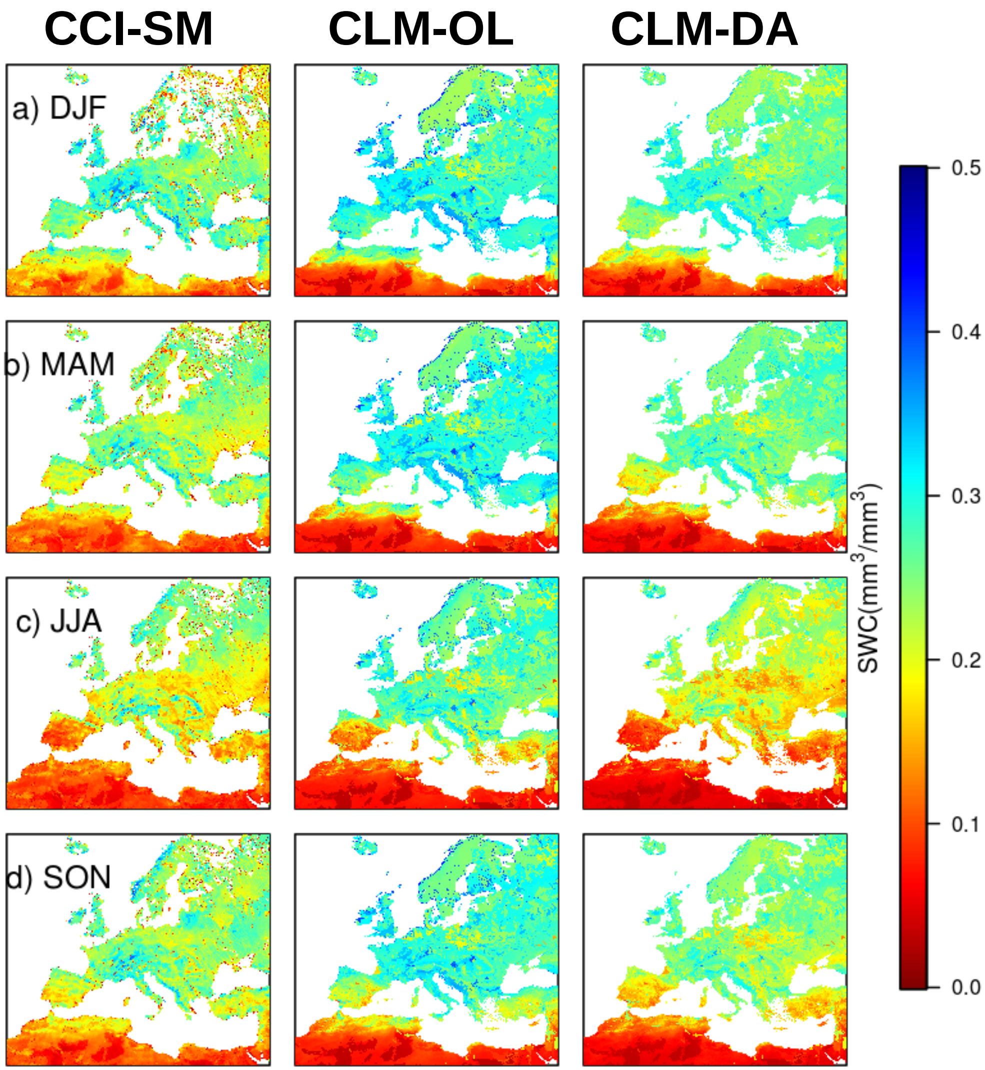
PDAF

25 km ESA CCI

3km ESSMRA

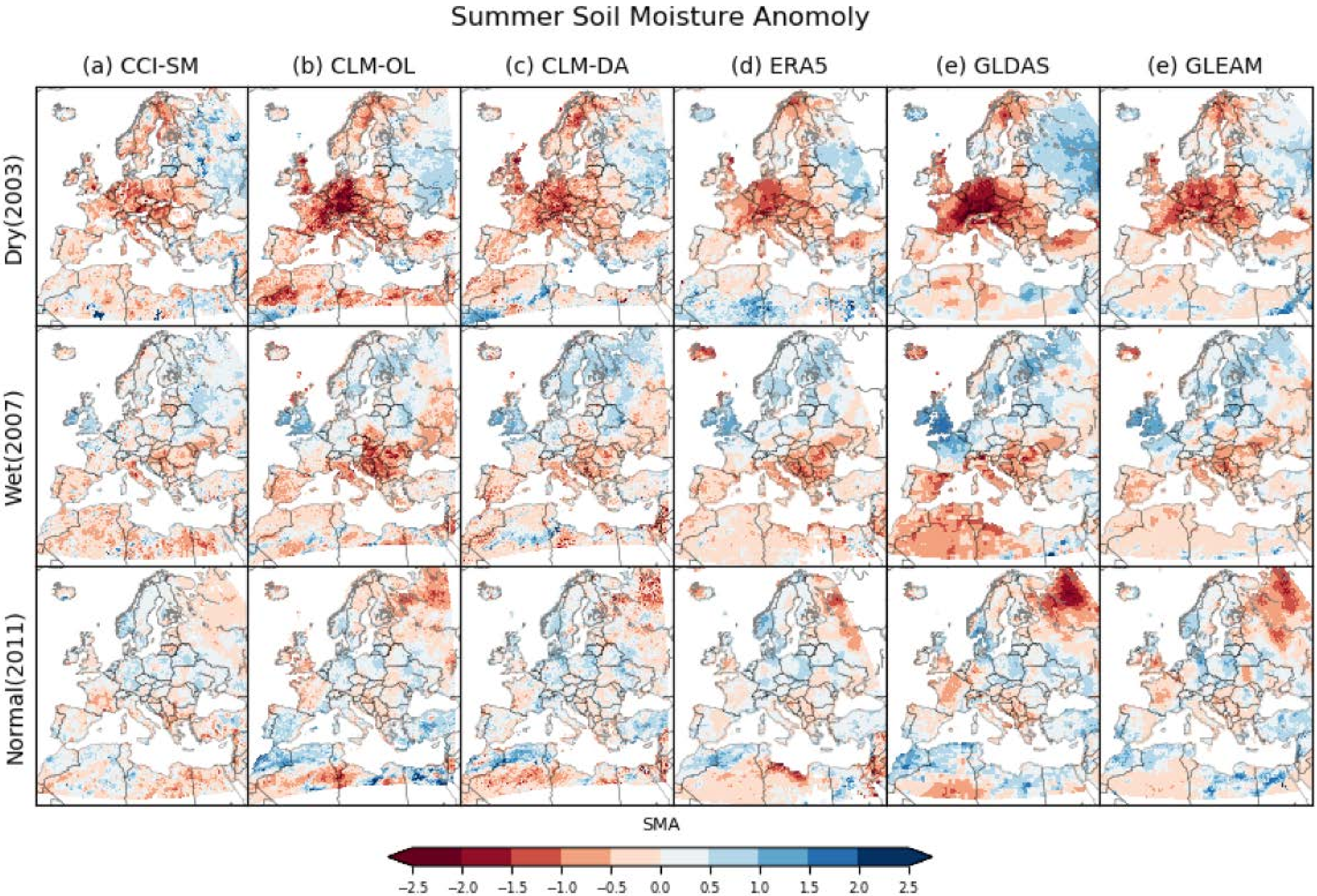
TSMP-PDAF EUROCORDEX DOMAIN

- Assimilation of coarse scale soil moisture (SM) in high resolution LSM CLM v. 3.5 (PDAF)
- 3km model resolution
- Model forcing COSMO-REA6 reanalysis (6km)
- SM product ESA CCI (25km)
- 100 grid cells randomly selected and used in DA
- Evaluation of OL simulations and DA-runs at all locations
- Comparison with gridded monthly runoff data E-RUN v1 and GRACE total water storage
- Improvements especially in summer and autumn



Naz et al., 2019, HESS

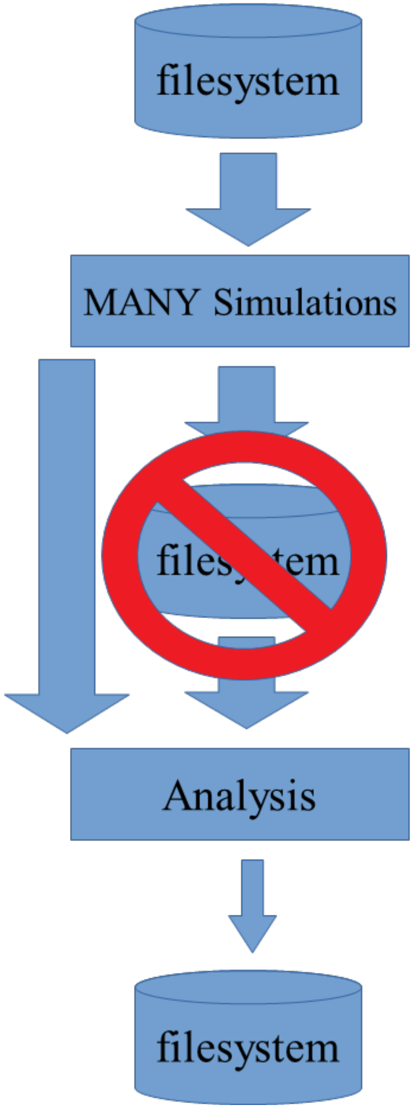
DA EUROCORDEX 2001-2015; EXAMPLES



Naz et al., 2020, Scientific Data

DEVELOPMENT: TSMP-PDAF-MELISSA

Traditional:



filesystem

MANY Simulations

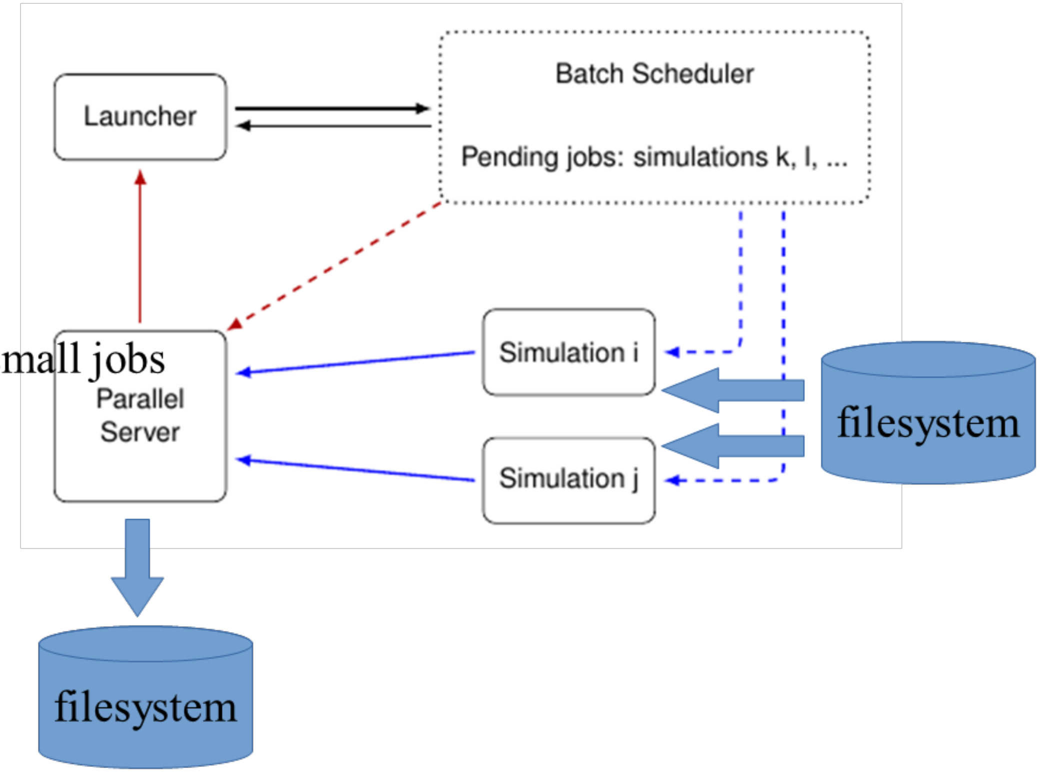
filesystem

Analysis

filesystem

MELISSA:

& Submit many small jobs



Launcher

Batch Scheduler

Pending jobs: simulations k, l, ...

Parallel Server

Simulation i

Simulation j

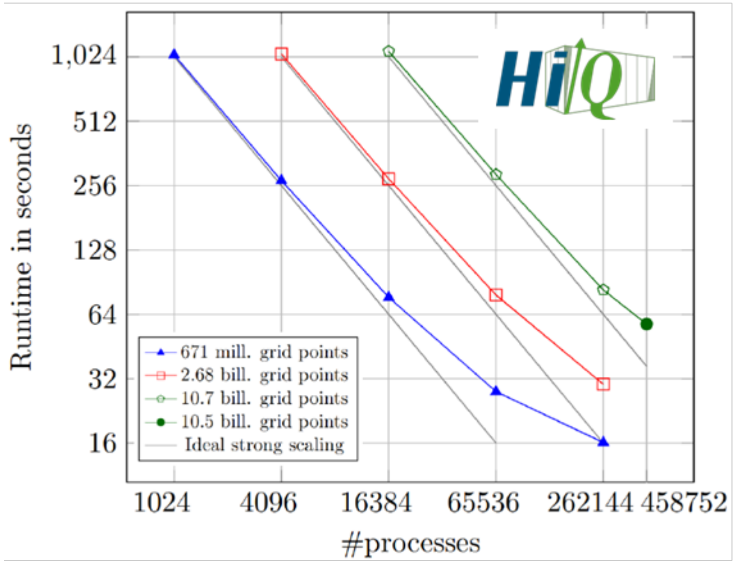
filesystem

filesystem

Work by Sebastian Friedman, INRIA, France, EU H2020 EoCoE

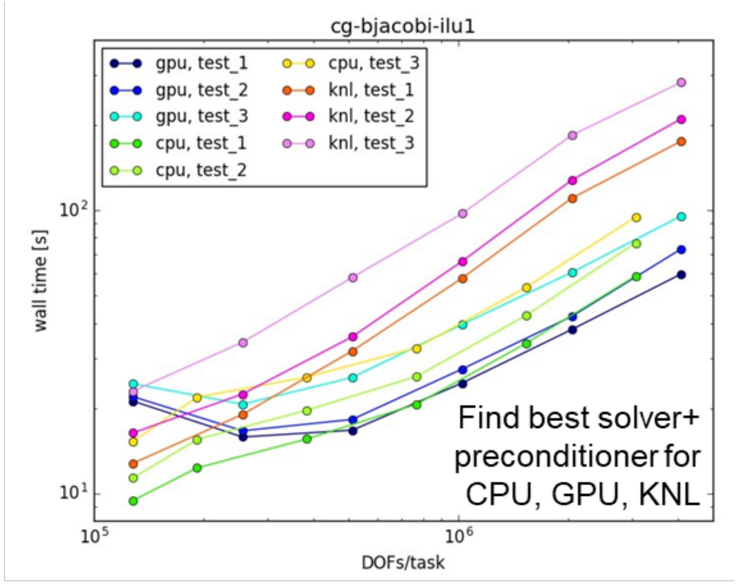
DEVELOPMENT: RUNNING TSMP ON GPU´S

Porting ParFlow (JURECA, Xeon Phi; JUWELS, NVIDIA GPUs)



Burstedde et al. (2017, arXiv), ParFlow+p4est (AMR)

Extreme scaling ParFlow model on JUQUEEN



Find best solver+ preconditioner for CPU, GPU, KNL

Python Mini-App testing of PETSc solver (permeability and input pressure matrix)

DEVELOPMENT: RUNNING TSMP ON GPU´S

- **Approaches**
  - Basis: Extensive performance profiling
  - CUDA, OpenCL
  - Domain specific language (Kokkos, RAJA Performance Portability Layer)
  - Optimized parallel I/O w/ netCDF

- **Testing and optimization ongoing**
  - In ParFlow, many loops/code regions parallelized with CUDA
  - CUDA unified memory implemented
  - Implementation of GPU linear solvers from EoCoE partners

- **Ongoing**
  - Workflows, big data capabilities

CONCLUSIONS AND OUTLOOK

- Terrestrial Systems Modelling is affected by large prediction uncertainties and needs large amounts of data to better constrain model predictions

- TSMP-PDAF is a highly efficient DA-framework which can assimilate observations from subsurface, land surface and atmosphere

- Applications from small catchment scale to continental scale

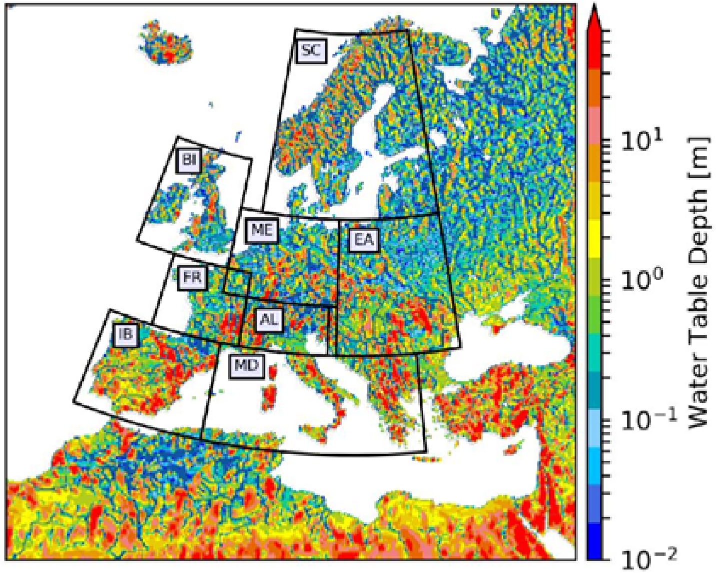
- Simultaneous assimilation of data from all three compartments is still pending (weakly coupled and fully coupled DA) → FOR2131

- Given extreme compute requirements only EnKF (and variants) were tested with TSMP-PDAF (~ 10^2 model runs).
- In smaller projects (groundwater model) other algorithms were tested which need up to 10^5 - 10^6 model evaluations, but are more accurate

SOME HPSC ACTIVITIES OF IBG-3 AT JSC

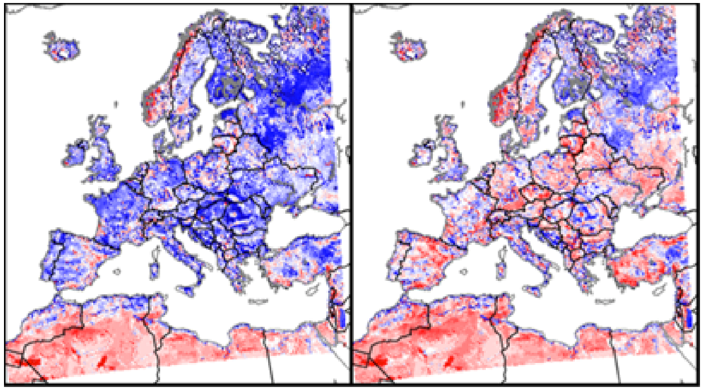
Water cycle and water resources modelling (G2A)

- Fully coupled TSMP
- Standalone CLM and ParFlow
- Sensitivity studies
- Climate projections
- Ecosystem reanalysis
- Forecasts and monitoring
- Convection perm. and LES

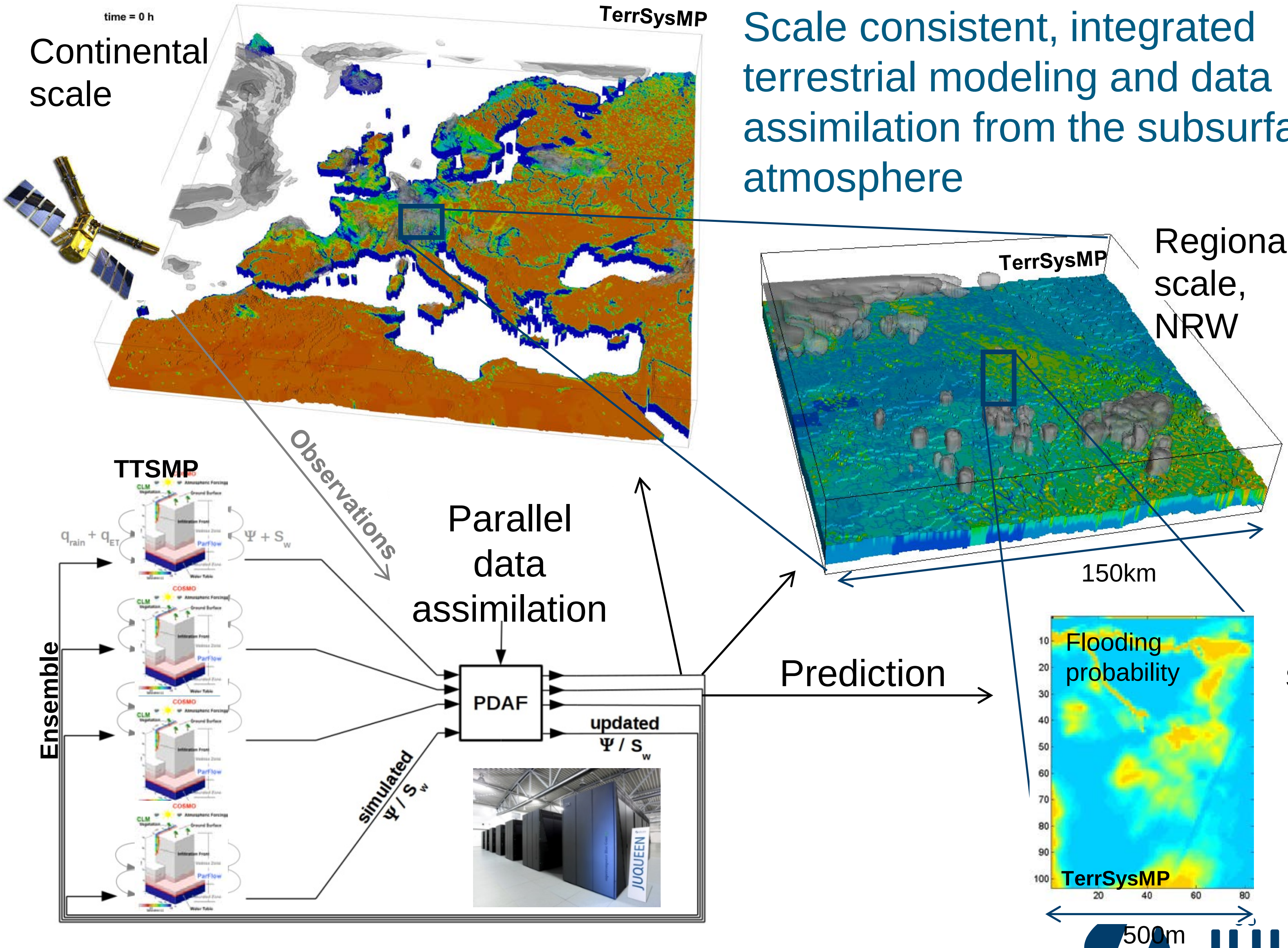


Data assimilation: Less uncertainties, better forecasts

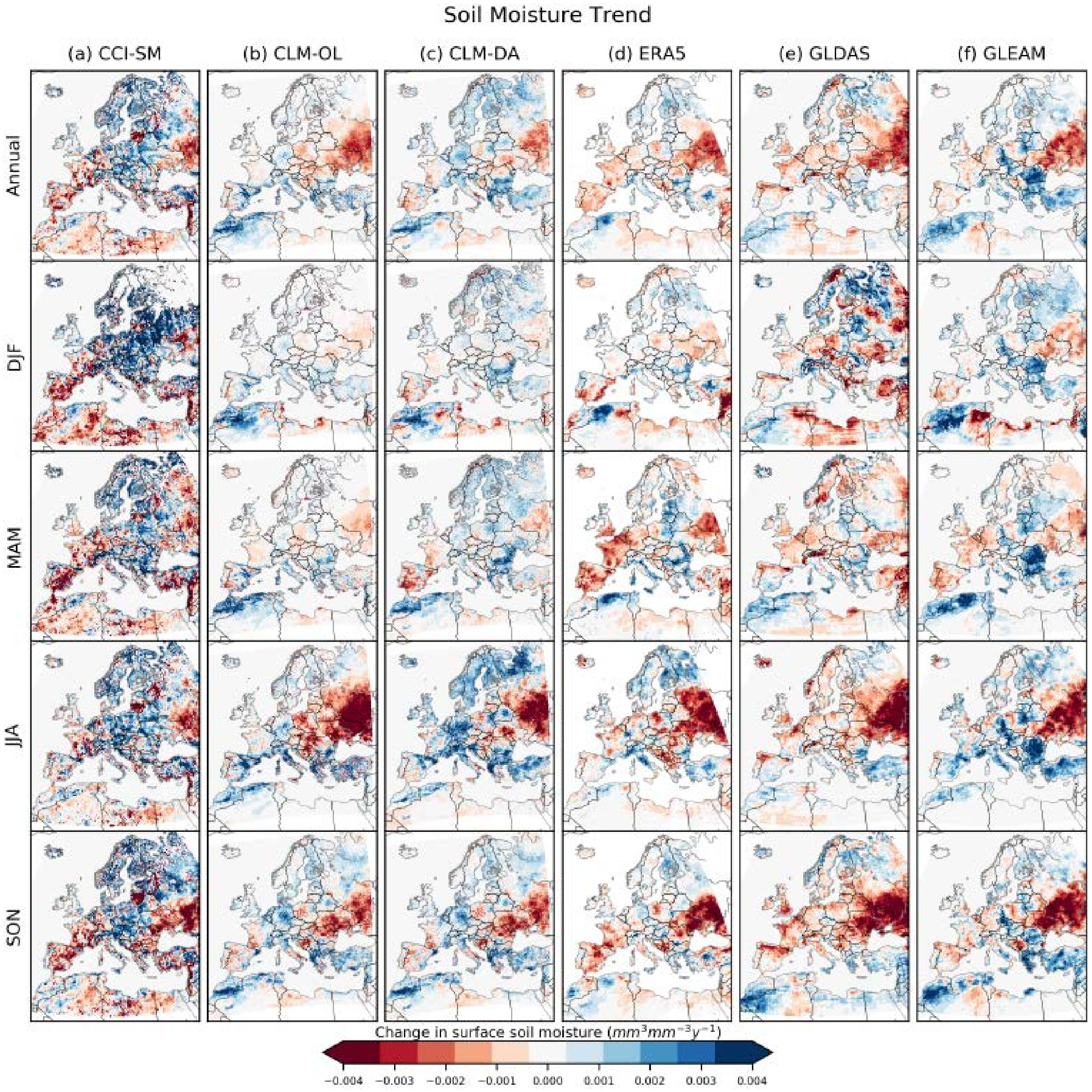
- Non-, weakly-, (fully-coupled)
- Hydrological and plant physiological quantities for state and parameter updates
- TSMP-PDAF, all in-memory
- Excellent scalability
- Research on DA algorithms



Scale consistent, integrated terrestrial modeling and data assimilation from the subsurface into atmosphere



DA EUROCORDEX 2001-2015; TRENDS



JÜLICH
Forschungszentrum