Data Assimilation with the Integrated Terrestrial System Model TSMP-PDAF

H.-J. Hendricks Franssen, W. Kurtz, B. Naz, S. Gebler

published in

NIC Symposium 2020
M. Müller, K. Binder, A. Trautmann (Editors)

Forschungszentrum Jülich GmbH,
John von Neumann Institute for Computing (NIC),
Schriften des Forschungszentrums Jülich, NIC Series, Vol. 50,
ISBN 978-3-95806-443-0, pp. 301.
http://hdl.handle.net/2128/24435

Data Assimilation with the Integrated Terrestrial System Model TSMP-PDAF

Harrie-Jan Hendricks Franssen[1], Wolfgang Kurtz[2], Bibi Naz[1], and Sebastian Gebler[1]

[1] IBG-3 Agrosphere, Forschungszentrum Jülich, 52425 Jülich, Germany
E-mail: {h.hendricks-franssen, b.naz, s.gebler}@fz-juelich.de

[2] Leibniz Supercomputing Centre, Boltzmannstraße 1, 85748 Garching, Germany
E-mail: wolfgang.kurtz@lrz.de

This paper discusses predictions with integrated terrestrial system models, which model water and energy cycles from the deep subsurface to the upper atmosphere. Such predictions can be improved by data assimilation, making use of supercomputing. Two examples on very different scales (hillslope and continental) illustrate the strengths and remaining uncertainties of the approach.

1 Introduction

Traditionally, in hydrology, soil science and ecology compartment-specific simulation models are employed. For example, groundwater models were developed able to simulate groundwater flow in (heterogeneous) aquifers, and soil hydrological models able to simulate water flow in the unsaturated zone including soil evaporation and plant transpiration. In the last two decades integrated hydrological modelling and integrated terrestrial systems modelling have gained importance. In these integrated models multiple compartments of the terrestrial system are simulated jointly, including their two-way feedbacks. The development of integrated models is driven by the fact that the different compartments of the terrestrial system show strong non-linear interactions. In addition, increase in compute power made integrated model simulation with many unknowns feasible, first on large computing clusters, and later also on PC's. Examples for integrated hydrological models, which simulate water flow in soil, aquifer and streams in a coupled fashion are ParFlow[1] and HydroGeoSphere.[2]

Land surface models simulate the exchange of water and energy – the most recent model generation(s) also carbon and nitrogen – between the land and the atmosphere. Land surface models incorporated over time increasingly sophisticated representations of soil and vegetation to better simulate these exchange processes. Land surface models were therefore in an early stage already integrated models. Nevertheless, land surface models required a more mechanistic based representation of subsurface flow and heat transport, which gave rise to the development of integrated terrestrial system models, which couple land surface models and integrated hydrological models, like for example CLM-ParFlow.[3] More recently, such models were also coupled to atmospheric circulation models, as is the case of the Terrestrial System Modelling Platform (TSMP) that we use in this work.[4]

Although models like TSMP provide a more mechanistic and complete representation of the water and energy cycles in terrestrial systems than classical compartment-specific models, simulations are affected by large uncertainties related to uncertain initial and boundary conditions and input parameters. Data assimilation approaches allow to estimate and reduce simulation uncertainties. Data assimilation can constrain initial conditions and parameters by correcting model simulations with measurements, which is based on the optimal weighting of simulated values on the one hand and measured values on the other hand. Ensemble based data assimilation, which characterises model uncertainty with a larger number of model runs, is used in this work. Examples of ensemble-based data assimilation algorithms are the Ensemble Kalman Filter[5] and the Particle Filter.[6]

This paper discusses first the Terrestrial Systems Modelling Platform, and the data assimilation framework coupled to it. Afterwards, two application examples are presented.

2 The Terrestrial Systems Modelling Platform TSMP

The Terrestrial System Modelling Platform (TSMP)[4] is a modular Earth system model which combines already pre-existing parallel compartment models for the atmosphere (COSMO[7]), the land surface (CLM version 3.5[8]) and the subsurface (ParFlow[1, 9, 10]). COSMO is a convection-permitting atmospheric model which is used as the operational forecast model of the German weather service. CLM is a land surface model that simulates the transfer and partitioning of energy, momentum, water and carbon fluxes between the atmosphere, vegetation and the subsurface. ParFlow calculates variably-saturated subsurface flow and surface water routing (using the kinematic wave approximation) in an integrated approach.[1]

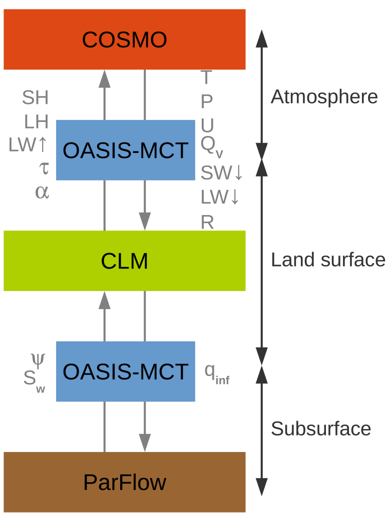

Figure 1. Coupling of the TSMP component models ParFlow (subsurface), CLM (land surface) and COSMO (atmosphere) by OASIS-MCT. The exchanged fluxes and state variables are: ψ (subsurface pressure), S_w (subsurface saturation), q_{inf} (net infiltration flux), SH (sensible heat flux), LH (latent heat flux), LW $\uparrow$ (outgoing long wave radiation), τ (momentum flux), α (albedo), P (air pressure), T (air temperature), U (wind velocity), SW $\downarrow$ (incoming short wave radiation), LW $\downarrow$ (incoming long wave radiation), Q_{V} (specific humidity) and R (precipitation).

The coupling library OASIS-MCT[11, 12] is used to connect the three component models of TSMP by exchanging information on fluxes and state variables at the conceptual boundaries of the respective compartment models (see Fig. 1) meaning that each compartment model acts as a lower or upper boundary condition for the adjacent compartment model. The data exchange between models via OASIS-MCT is managed by one global communicator (`MPI_COMM_WORLD`) that is shared between the involved compartment models. Information on exchanged variables and corresponding interpolation and scaling options as well as the required communication patterns are defined in the initialisation phase of the compartment models via OASIS-MCT library calls. The models are then dynamically coupled at runtime through information exchange at predefined time intervals. Key features of TSMP are its integral view on terrestrial processes, its improved physical process description (as compared to single compartment models), its modularity (different model combinations can be chosen) and a good scalability on HPC platforms.[12]

3 The Data Assimilation Framework TSMP-PDAF

Data assimilation with TSMP is enabled by coupling TSMP to the Parallel Data Assimilation Framework (PDAF).[13] PDAF[14] is a generic data assimilation library that provides a variety of parallel implementations of ensemble filters that are commonly used for data assimilation. In addition, PDAF provides functionality for data exchange between ensemble members and the parallel data assimilation algorithms. For the coupling of TSMP with PDAF, a set of functions needed to be defined, that provide PDAF with information and data for the assimilation. This includes, for example, functions to extract the relevant state variables of the TSMP models and functions for providing information on the observation data.

In the initialisation phase of TSMP-PDAF (before TSMP models are initialised), PDAF defines three MPI communicators for facilitating data exchange for the assimilation and the ensemble propagation (Fig. 2). The "model communicator" is defined for each model realisation and is used for the model forward integration. The "filter communicator" is used for the parallel filter algorithm and is only active for the processors of the first model realisation. The "coupling communicator" is used to collect/distribute information on model states across different model realisations to/from the filter communicator. Note that each rank in the model and coupling communicators usually only holds a part of the global model state vector. For using PDAF with the coupled TSMP model, modifications of the OASIS-MCT library were necessary to allow the ensemble integration. Specifically, for the data exchange between different compartment models of TSMP with OASIS-MCT, the model communicators are used instead of `MPI_COMM_WORLD` (see Sec. 2).

The PDAF filter algorithms are called during the time integration of the TSMP models at predefined assimilation intervals when measurements become available. In a first step, the relevant model states of the ensemble members are collected to the filter communicator by the coupling communicator. Then the filter algorithm is executed within the filter communicator using the available observation data. The updated model states are then distributed to the ensemble members via the coupling communicator and the TSMP models are integrated to the next assimilation step.

Similarly to TSMP, the data assimilation framework TSMP-PDAF is modular and thus allows to perform data assimilation with different model combinations as well as stand-

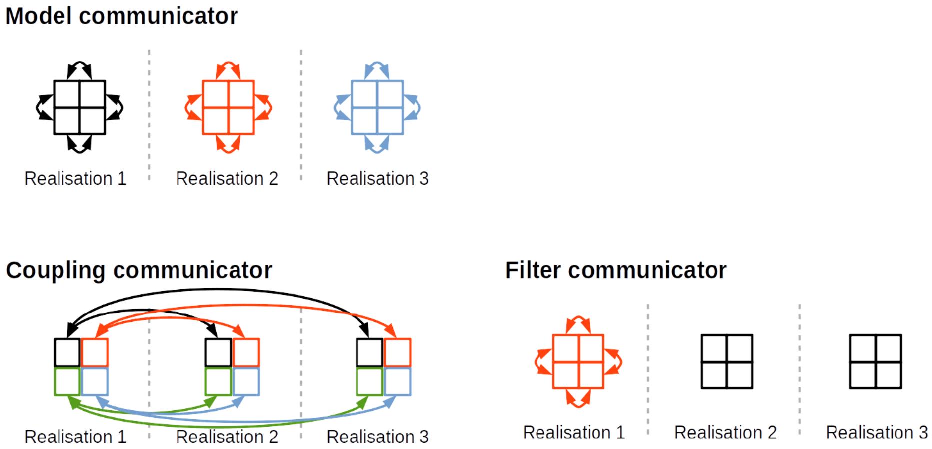

Figure 2. Communicators in PDAF for a parallel setup with 3 ensemble members and 4 processors per ensemble member. Colours indicate the membership of the respective processors and arrows exemplify the parallel communication between the different processors.

alone models of TSMP. Currently supported observation data are point-scale soil moisture measurements (for ParFlow and CLM), river stages and groundwater levels (both for ParFlow). TSMP-PDAF also allows to update several model parameters (hydraulic conductivity, Manning's coefficients and soil texture) with the mentioned observations.

The scalability of TSMP-PDAF was tested with a synthetic data assimilation experiment with a ParFlow-CLM model configuration on JUQUEEN (Fig. 3) using different scenarios regarding model output and ensemble quality. An ideal setup with identical ensemble members was used to determine the computational overhead of the ensemble integration and the filtering step. Results showed that the overhead starts to become relevant with more than 8192 processors. When using a realistic ensemble with varying model parameters, the scalability decreases due to the load imbalance caused by different run times of the individual ensemble members. The model output scenarios showed low variations of the total run time for the realistic ensemble meaning that the load imbalance had a more prominent effect for the utilised data assimilation setup.

4 Example of Data Assimilation for Soil Moisture

4.1 Hillslope Scale High-Resolution Integrated Land Surface-Subsurface Modelling

The CLM and ParFlow components of the TSMP-model were used for land surface-subsurface modelling of the highly equipped Rollesbroich site (0.38 km^2) in the Eifel hills in North-Rhine-Westphalia in Germany. At this site, information from a soil moisture sensor network measuring soil moisture content at 61 locations (5, 20 and 50 cm depth) was

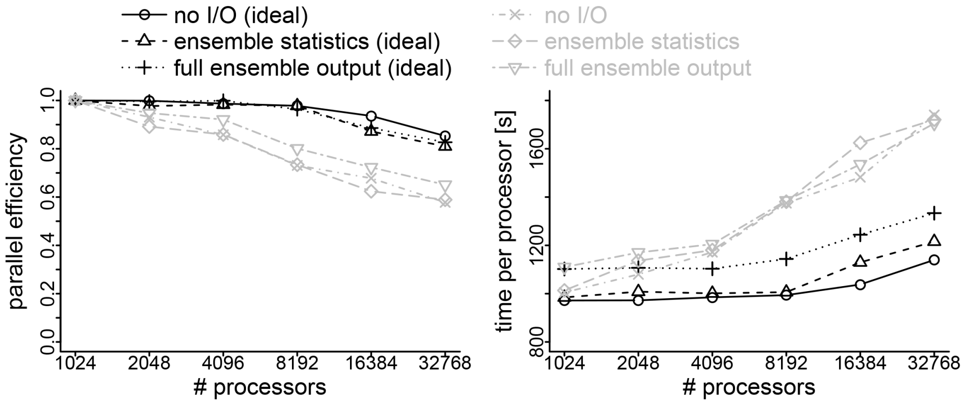

Figure 3. Scaling behaviour (left) and timing information (right) for TSMP-PDAF for a weak scaling test on JUQUEEN. Black lines show results for an idealised test case (identical ensemble members) and grey lines show results for a heterogeneous ensemble. The number of ensemble members is increased from 8 to 256. Each ensemble member used 32 processors for CLM and 96 processors for ParFlow. Different lines refer to experiments with different model output.

available. The research question was, whether for this hillslope research site the combination of physically based modelling with the TSMP-model and the dense observation network, allow to make an accurate characterisation of hydrological states and fluxes including discharge. In other works, it is often found that predicting discharge at the hillslope scale is particularly challenging.

This question was addressed with TSMP-PDAF and for the year 2011, where the Ensemble Kalman Filter was used for assimilating the data from the soil moisture sensor network. An ensemble of 128 (or 256) realisations with different 3D-heterogeneous fields of Mualem-van Genuchten parameters was generated. Also atmospheric forcings were considered uncertain and perturbed for each of the 128 (or 256) realisations. Different assimilation scenarios were run, including state updating alone, and joint state and parameter updating. Simulations were also carried out with a synthetic test case mimicking the Rollesbroich site, to get more insight in the role of model structural errors.

The combination of joint updating of model states and hydraulic conductivity was more efficient for soil water content (SWC) characterisation than state updating alone for the real-world case. On average, the root mean square error (RMSE), which measures the difference between measured and simulated SWC at the sensor locations, was reduced by 14% if states and parameters were updated jointly, but discharge estimation was not improved significantly. See also Fig. 4. Synthetic simulations showed much better results with an overall RMSE reduction by 55% at independent verification locations in case of daily SWC data assimilation including parameter estimation. Individual synthetic data assimilation scenarios with parameter estimation showed an increase of the Nash-Sutcliffe-Efficiency for discharge from -0.04 for the open loop run to 0.61. This shows that data assimilation in combination with high-resolution physically-based models can potentially strongly improve soil moisture and discharge estimation at the hillslope scale. Additional simulation experiments were performed to understand the difference between the real-world

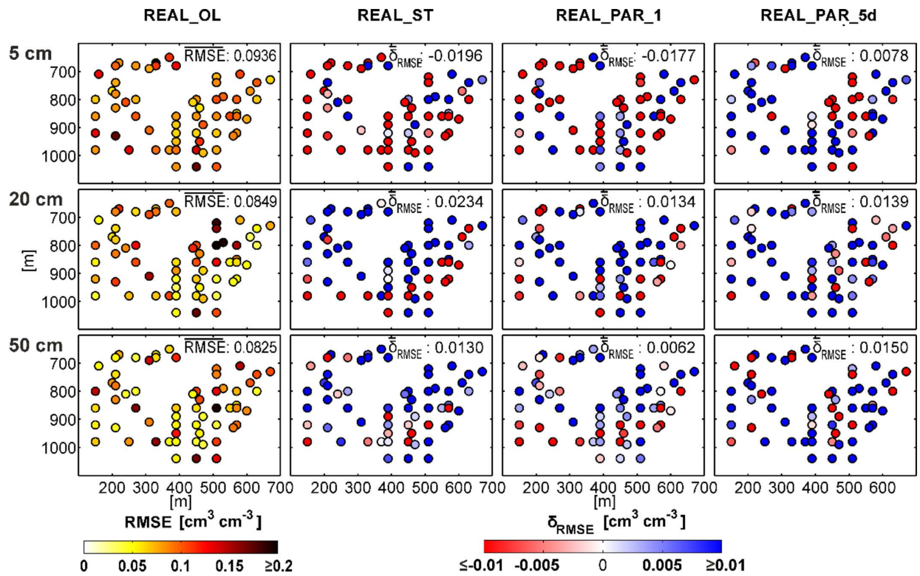

Figure 4. RMSE of SWC at individual locations for the open loop runs (left column) and changes in RMSE (increase implies improvement and decrease implies impoverishment) for three data assimilation scenarios (state updating alone (ST), daily joint state and parameter updating (PAR1) and joint state and parameter updating every five days (PAR5d) (three columns on the right)) of the real-world case for 2011.

case and the synthetic case. It was found that erroneous prior values of the geostatistical parameters are for example already able to explain a considerable part of the difference in performance between the real-world and synthetic case. On the other hand, if only the saturated hydraulic conductivity was unknown (and other soil hydraulic parameters were known), the performance hardly improved compared to the case that all soil hydraulic parameters were unknown. In summary, the large performance difference between synthetic and real-world experiments indicates the limits of such an approach associated with model structural errors like errors in the prior geostatistical parameters.

4.2 Continental-Scale High-Resolution Land Surface Data Assimilation System

Soil moisture is a key state variable which controls the exchange of water, energy and carbon fluxes between the land surface and atmosphere.[15] As a result, it plays an important role in many regional–scale applications, including meteorology, hydrology, flood forecasting, drought monitoring, agriculture and climate change impact studies.[16] Because of its high spatiotemporal variability, it is difficult to monitor soil moisture at large spatial scales. The knowledge of soil moisture at large scale, with reasonable temporal and spatial resolution, is therefore needed to provide locally representative information of soil moisture for regional hydrologic and agriculture applications. However, soil moisture is a difficult variable to obtain because there are no high-resolution soil moisture observations available at the continental scale, and observations from measurements or remote sensing products are sparse and temporally and spatially discontinuous. While land surface models can provide high resolution large scale high resolution soil moisture estimates with complete spatial and temporal coverage, the soil moisture fields from models often suffer substantial errors owing to errors either in model forcing (such as precipitation, temperature and radiation) or inadequate model physics. We used the land surface data assimilation system CLM-PDAF to utilise the coarse resolution satellite soil moisture data to update the soil moisture estimates from the land surface model. A key capability of CLM-PDAF is the support for data assimilation that combines land surface processes with satellite and in-situ observations for the estimation of optimal land surface states. The data assimilation structure in CLM-PDAF allows to directly ingest remotely sensed high resolution observations of land surface conditions to produce accurate, spatially and temporally consistent fields of land surface states, with reduced associated error. The CLM-PDAF uses an Ensemble Kalman Filter (EnKF) algorithm to generate assimilated or reanalysis products. To effectively simulate the background-error covariances, a large enough ensemble size needs to be maintained in the data assimilation process, which linearly increases the computational resource requirements. Hence, we implemented the CLM-PDAF over Europe to provide downscaled estimates of the soil moisture with complete spatiotemporal coverage by combining historical satellite soil moisture (SM) observations with a high resolution land surface model (LSM) using data assimilation techniques. Using the CLM-PDAF, the satellite based soil moisture dataset ESA CCI (the European Space Agency Climate Change Initiative[17]) was assimilated into CLM using the EnKF[5] producing a high-resolution European surface soil moisture (SSM) reanalysis (called ESSMRA hereafter) dataset. This product overcomes the shortcomings of sparse spatial and temporal datasets and provides a better estimate of SM than obtained only by modelling or by sparse observations alone.

4.2.1 Work Flow for Generation of ESSMRA

The 3 km ESSMRA is generated by first implementing the regional land surface model setup coupled with the data assimilation framework as shown in Fig. 5a. In the second step the ESA CCI satellite-based data is assimilated into the CLM-PDAF setup to generate the daily 3 km ESSMRA product over Europe for the 2000 – 2015 time period (Fig. 5b). The ESSMRA dataset is also compared with other global soil moisture reanalysis products from the European Centre for Medium-Range Weather Forecasts Reanalysis 5[18](ERA5), the Global Land Data Assimilation System[19] (GLDAS) and the Global Land Evaporation Amsterdam Model[20] (GLEAM) data available at $0.25°$ resolution and hourly temporal resolution.

To assess the ability of ESSMRA to capture short term soil moisture variability in comparison to other existing reanalysis products (such as ERA5, GLDAS and GLEAM), summer standardised anomalies of SM were calculated ($SMA = \frac{SM_{jja} - \overline{SM}}{\sigma_{SM}}$) for a dry year (2003), wet year (2007) and normal year (2011) using average SM values over June, July and August (JJA) relative to the mean JJA SM for the 2000–2015 period. The spatial distribution shows patterns of positive and negative SM anomalies over Europe across all datasets for dry, wet and normal years (Fig. 6). For dry year 2003 (a record heat wave over Europe), CLM-DA (CLM estimated soil moisture with data assimilation) shows a similar area extent of negative anomalies as the ESACCI dataset and ERA5, whereas CLM-OL (CLM estimated soil moisture without data assimilation), GLDAS and GLEAM exhibits much stronger negative anomalies over central Europe. The SM anomaly from CLM-DA for the wet and normal years (2007, 2011) has a better match with ESACCI and other

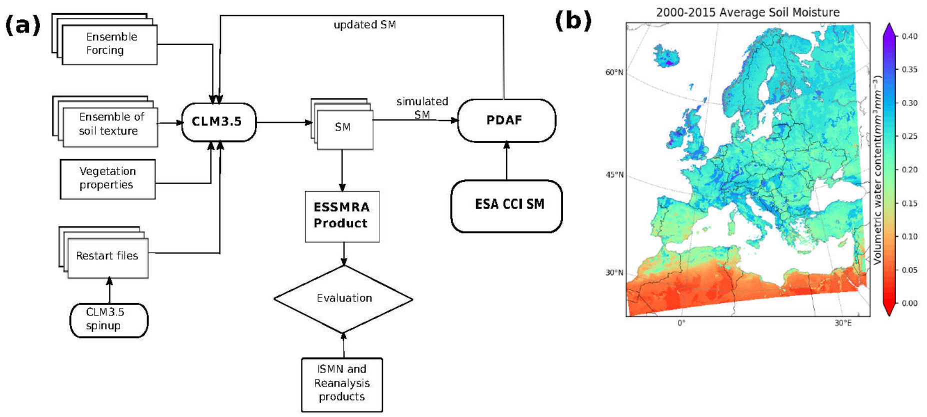

Figure 5. (a) Schematic of CLM-PDAF workflow adopted to generate high resolution ESSMRA product. (b) Average soil moisture (mm^3 mm^{-3}) for 2000 – 2015 time period.

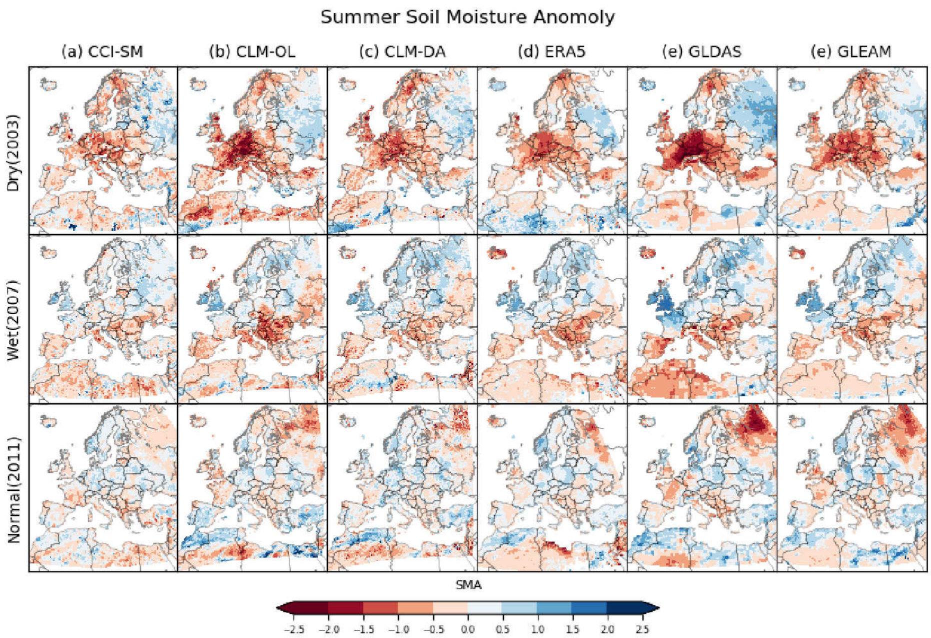

Figure 6. Spatial distribution of the standardised summer (JJA) soil moisture anomaly for the year 2003 (dry year), 2007 (wet year) and 2011 (normal year) and compared with existing reanalysis products. The summer anomaly is calculated for (a) satellite (ESACCI), (b) CLM-OL (c) CLM-DA, (d) ERA5, (e) GLDAS and (f) GLEAM.

reanalysis datasets except GLDAS which shows much stronger wet and dry anomalies than others.

5 Concluding Remarks

The parallel data assimilation framework (PDAF) has been coupled to the terrestrial modelling platform (TSMP) in order to condition integrated terrestrial system simulations to measurements. The framework shows a very good scalability on the Jülich supercomputers. The performance of TSMP-PDAF has been tested in synthetic and real-world cases, and at different spatial scales. Here we illustrate that the data assimilation performed well for simulations at the hillslope scale and continental scale.

Acknowledgements

The authors gratefully acknowledge the computing time granted through JARA, GCS and ESM on the supercomputers JURECA, JUQUEEN and JUWELS at Jülich Supercomuting Centre (JSC).

References

1. S. J. Kollet and R. M. Maxwell, *Integrated surface-groundwater flow modeling: A free-surface overland flow boundary condition in a parallel groundwater flow model*, Advances in Water Resources **29**, 945–958, 2006.
2. R. Therrien and E. Sudicky, *Three-dimensional analysis of variably-saturated flow and solute transport in discretely-fractured porous media*, Journal of Contaminant Hydrology **23**, 1–44, 1996.
3. S. Kollet and R. M. Maxwell, *Capturing the influence of groundwater dynamics on land surface processes using an integrated, distributed watershed model*, Water Resources Research **44**, W02402, 2008.
4. P. Shrestha, M. Sulis, M. Masbou, S. Kollet, and C. Simmer, *A Scale-Consistent Terrestrial Systems Modeling Platform Based on COSMO, CLM, and ParFlow*, Monthly Weather Review **142**, 3466–3483, 2014.
5. G. Burgers, P. J. van Leeuwen, and G. Evensen, *Analysis scheme in the ensemble Kalman filter*, Monthly Weather Review **126**, 1719–1724, 1998.
6. N. J. Gordon, D. J. Salmond, and A. F. M. Smith, *Novel-approach to nonlinear non-Gaussian bayesian state estimation*, IEEE Proceedings-F Radar and Signal Processing **140**, 107–113, 1993.
7. M. Baldauf, A. Seifert, J. Förstner, D. Majewski, M. Raschendorfer, and T. Reinhardt, *Operational Convective-Scale Numerical Weather Prediction with the COSMO Model: Description and Sensitivities*, Monthly Weather Review **139**, 3887–3905, 2011.
8. K. W. Oleson, G.-Y. Niu, Z.-L. Yang, D. M. Lawrence, P. E. Thornton, P. J. Lawrence, R. Stöckli, R. E. Dickinson, G. B. Bonan, S. Levis *et al.*, *Improvements to the Community Land Model and their impact on the hydrological cycle*, Journal of Geophysical Research **113**, G01021, 2008.
9. S. F. Ashby and R. D. Falgout, *A parallel multigrid preconditioned conjugate gradient algorithm for groundwater flow simulations*, Nuclear Science and Engineering **124**, 145–159, 1996.

10. R. M. Maxwell, *A terrain-following grid transform and preconditioner for parallel, large-scale, integrated hydrologic modeling*, Advances in Water Resources **53**, 109–117, 2013.
11. S. Valcke, *The OASIS3 coupler: a European climate modelling community software*, Geoscientific Model Development **6**, 373–388, 2013.
12. F. Gasper, K. Goergen, P. Shrestha, M. Sulis, J. Rihani, M. Geimer, and S. Kollet, *Implementation and scaling of the fully coupled Terrestrial Systems Modeling Platform (TerrSysMP v1.0) in a massively parallel supercomputing environment – a case study on JUQUEEN (IBM Blue Gene/Q)*, Geoscientific Model Development **7**, 2531–2543, 2014.
13. W. Kurtz, G. He, S. J. Kollet, R. M. Maxwell, H. Vereecken, and H. J. Hendricks Franssen, *TerrSysMP-PDAF (version 1.0): a modular high-performance data assimilation framework for an integrated land surface-subsurface model*, Geoscientific Model Development **9**, 1341–1360, 2016.
14. L. Nerger and W. Hiller, *Software for ensemble-based data assimilation systems-Implementation strategies and scalability*, Computers & Geosciences **55**, 110–118, 2013.
15. S. I. Seneviratne, T. Corti, E. L. Davin, M. Hirschi, E. B. Jaeger, I. Lehner, B. Orlowsky, and A. J. Teuling, *Investigating soil moisture-climate interactions in a changing climate: A review*, Earth-Science Reviews **99**, 125–161, 2010.
16. L. Brocca, F. Melone, T. Moramarco, and R. Morbidelli, *Spatial-temporal variability of soil moisture and its estimation across scales*, Water Resources Research **46**, W02516, 2010.
17. W. Wagner, W. Dorigo, R. de Jeu, D. Fernandez, J. Benveniste, E. Haas, and M. Ertl, *Fusion of active and passive microwave observations to create an essential climate variable data record on soil moisture*, ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS Annals) **7**, 315–321, 2012.
18. "Copernicus Climate Change Service (C3s) (2017): ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate, Copernicus Climate Change Service Climate Data Store (CDS)".
19. M. Rodell, P. R. Houser, U. Jambor, J. Gottschalck, K. Mitchell, C.-J. Meng, K. Arsenault, B. Cosgrove, J. Radakovich, M. Bosilovich, J. K. Entin, J. P. Walker, D. Lohmann, and D. Toll, *The Global Land Data Assimilation System*, Bulletin of the American Meteorological Society **85**, 381–394, 2004.
20. B. Martens, D. G. Miralles, H. Lievens, R. van der Schalie, R. A. M. de Jeu, D. Fernández-Prieto, H. E. Beck, W. A. Dorigo, and N. E. C. Verhoest, *GLEAM v3: satellite-based land evaporation and root-zone soil moisture*, Geoscientific Model Development **10**, 1903–1925, 2017.