https://doi.org/10.5194/gmd-2021-38
Preprint. Discussion started: 8 September 2021

Geoscientific Model Development Discussions

Coupling the Community Land Model version 5.0 to the parallel data assimilation framework PDAF: Description and applications

Lukas Strebel[1,2], Heye R. Bogena[1,2], Harry Vereecken [1,2], Harrie-Jan Hendricks Franssen [1,2]

[1] Agrosphere Institute, IBG-3, Forschungszentrum Jülich GmbH, Germany
[2] Centre for High-Performance Scientific Computing in Terrestrial Systems: HPSC TerrSys, Geoverbund ABC/J, Leo-Brandt-Strasse, 52425 Jülich,

Correspondence to: Lukas Strebel (l.strebel@fz-juelich.de)

Abstract. Land surface models are important for improving our understanding of the earth system. They are continuously improving and becoming more accurate in describing the varied surface processes, e.g. the Community Land Model version 5 (CLM5). Similarly, observational networks and remote sensing operations are increasingly providing more and higher quality data. For the optimal combination of land surface models and observation data, data assimilation techniques have been developed in the past decades that incorporate observations to update modeled states and parameters. The Parallel Data Assimilation Framework (PDAF) is a software environment that enables ensemble data assimilation and simplifies the implementation of data assimilation systems in numerical models. In this paper, we present the further development of the PDAF to enable its application in combination with CLM5. This novel coupling adapts the optional CLM5 ensemble mode to enable integration of PDAF filter routines while keeping changes to the pre-existing parallel communication infrastructure to a minimum. Soil water content observations from an extensive in-situ measurement network in the Wüstebach catchment in Germany are used to illustrate the application of the coupled CLM5+PDAF system. The results show overall reductions in root mean square error of soil water content from 7% up to 35% compared to simulations without data assimilation. We expect the coupled CLM5+PDAF system to provide a basis for improved regional to global land surface modelling by enabling the assimilation of globally available observational data.

1 Introduction

The land surface forms the interface between the atmosphere and the lithosphere and plays a crucial role in the global climate system. Therefore, land surface models (LSMs) are an important tool to progress our understanding of the Earth system. LSMs represent a wide variety of processes from energy partitioning and mass exchanges to hydrological and ecological processes. The research community has developed sophisticated parameterizations and combined them into increasingly complex and accurate LSMs (Overgaard et al., 2006). However, predictions with LSMs are still affected by various important sources of uncertainty, including initial conditions, parameters, parameterization (e.g. surface and subsurface water flow), and effects of the commonly used coarse resolution of LSMs (Wood et al., 2011). Therefore, observational data are often used to improve model predictions. Here we focus on soil water content (SWC) as it is a key variable that strongly influences the partitioning

https://doi.org/10.5194/gmd-2021-38
Preprint. Discussion started: 8 September 2021


Geoscientific Model Development Discussions

of latent and sensible heat flux as well as the partitioning of precipitation into surface runoff and infiltration (e.g. Vereecken et al. 2008). Furthermore, SWC has a strong influence on vegetation growth and modulates fire risks (e.g. Buotte et al. 2019). A common LSM is the Community Land Model (CLM) (Lawrence et al. 2019), of which the performance has already been evaluated in various studies with observational data. For example, Hudiburg et al. (2013) used a single-point setup of the CLM4.0 model to predict net and gross primary production of forested FLUXNET sites in Oregon, USA. Similar CLM single-point setups were also used to perform model sensitivity studies. For instance, Zhang et al. (2019) adjusted vegetation phenology parameters of the temperate grassland plant functional type in CLM4.5 to reduce an overestimation of growing-season LAI and annual gross primary production, while enhancing the partitioning of evapotranspiration for the study site. Similarly, Post et al. (2017) also used CLM4.5 single-point setups to estimate net carbon fluxes at four European sites and they improved the assessment of annual net ecosystem exchange by estimating ecosystem parameters using a Markov chain Monte Carlo method.

On the other hand, observational SWC data also face various limitations and uncertainties (Vereecken et al., 2008). For instance, high-quality in-situ SWC measurements usually only cover relatively small areas, while remote sensing observations give only indirect information about SWC for the upper few centimeters of the soil and with a coarse spatiotemporal resolution. Data assimilation aims at optimally merging model simulations and measurement data, according to statistical optimality principles, so that the uncertainty of the model simulations is reduced and the accuracy improved. Commonly numerical models are implemented without intrinsic data assimilation and external frameworks are used to perform data assimilation. Coupling to a framework instead of implementing data assimilation inside the numerical model provides many advantages. External frameworks are usually built for modularity and extendibility, i.e., these frameworks provide multiple different data assimilation methods and can be updated when new methods are developed. Additionally, frameworks are usually optimized for parallel computing. Between frameworks and sometimes within frameworks we can further distinguish between two different approaches for the coupling of models with external frameworks. In case of offline coupling, the framework wraps around the model and does not modify the model. This non-intrusive method uses the input, output and restart functionalities of the model to perform data assimilation. In contrast, the online coupling framework is incorporated into the model code, which allows to perform data assimilation in the main memory during simulation avoiding costly file input/output operations. The Data Assimilation Research Testbed (DART) (Anderson et al., 2009), which was originally developed for data assimilation with atmospheric models, is commonly used for offline coupled data assimilation. Recently some studies have shown its application in combination with CLM. For example, Zhang et al. (2014) assimilated satellite snow cover fraction data from MODIS (Moderate Resolution Imaging Spectroradiometer) into CLM4.0 using DART, which led to improved snow depth predictions. Ling et al. (2019) assimilated the Global Land Surface Satellite (GLASS) leaf area index (LAI) product into CLM4.0 using DART. They showed that updating both model LAI and leaf C/N can reduce the largest bias from $5m^2/m^2$ by $1m^2/m^2$ and significantly improve LAI predictions especially in forested regions. In another study, LAI and biomass observations were assimilated into a single-point CLM4.5 model for a semiarid ecosystem site in central New Mexico, USA, which improved the simulation of the carbon cycle (Fox et al. 2018). Recently, DART has been also been used to assimilate

https://doi.org/10.5194/gmd-2021-38
Preprint. Discussion started: 8 September 2021


Geoscientific Model Development Discussions

brightness temperature data from the Advanced Microwave Scanning Radiometer for Earth Observing System (AMSR-E) into CLM4.0 on a global scale to improve the prediction of soil water content (Zhao et al., 2016). In this study, it could be shown that soil water content simulation can be improved by data assimilation, but some of the systematic biases of CLM4 simulations could not be resolved. The Parallel Data Assimilation Framework (PDAF) (Nerger et al. 2005) has also been used in various studies to assimilate SWC measurements into different CLM model versions. In a recent study, PDAF was used to assimilate the ESA CCI microwave soil water content product in CLM3.5 with the ensemble Kalman filter to improve European predictions of soil water content and runoff estimations (Naz et al. 2019, 2020).

In this study, we choose PDAF as a framework for the data assimilation because it provides many data assimilation algorithms, supports online coupling, and includes templates for the modifications to the model code that are necessary for the coupling with CLM5. Additionally, PDAF is also part of the modular Terrestrial System Modeling Platform (TSMP) (Shrestha et al. 2014). PDAF has previously been coupled to CLM 3.5 within TSMP (Kurtz et al. 2016) and thus coupling PDAF to CLM5 has the potential benefit of simplifying future couplings to the other components of TSMP.

To illustrate the potential of the CLM5 PDAF coupling, we also present an application using the ensemble Kalman Filter to perform simultaneous state and parameter updates in the forest headwater catchment Wüstebach. The Wüstebach catchment is part of the TERENO network and various hydrological models have already been applied to it, e.g. HydroGeoSphere (Cornelissen et al., 2016; 2014); MIKE-SHE (Koch et al., 2016) and CLM-Parflow (Fang et al., 2015; 2016). Some of these modelling studies have focused on the spatial and temporal analysis of the effect of different parameterization approaches to represent the heterogeneous soil properties (Cornelissen et al., 2014; Fang et al. 2015; 2016). Koch et al. (2016) compared CLM-Parflow, HydroGeoSphere and MIKE-SHE and concluded that the consideration of heterogeneous porosities can increase model performance depending on the model structure. In contrast to these detailed distributed catchment studies, we model the study site from the viewpoint of a larger regional model where the catchment is represented by a single grid cell.

In this paper, we present the further development of the latest version of CLM (CLM5) to enable the use of PDAF and thus explore the potential of data assimilation in CLM5 and test the potential for updating model parameters. Furthermore, we investigate whether updating of the soil organic matter parameter via data assimilation can further improve the prediction of soil water with CLM5.

The paper is structured as follows: First, we give a short description of CLM5 and PDAF and then explain in detail how their coupling was realized. We then present the study site, the data used for the simulations, and the results for different data assimilation scenarios. The paper ends with a conclusion and outlook on further improvements that will be made, specifically concerning parameter updating.

https://doi.org/10.5194/gmd-2021-38
Preprint. Discussion started: 8 September 2021


Geoscientific Model Development Discussions

2 Methods

2.1 Model description

In this study, the Community Land Model 5.0 (CLM5) (Lawrence et al. 2019) is used to simulate land surface processes, in particular hydrological processes such as infiltration, evaporation from both soil and vegetation, transpiration, surface runoff and sub-surface drainage. We focus in particular on the simulation of the distribution and temporal dynamics of soil water within the soil column. Surface runoff is simulated in CLM5 using the SIMTOP model (Niu et al. 2005), which is based on the TOPMODEL approach (Beven and Kirkby 1979). Compared to previous versions, CLM5 allows a spatially variable soil depth with an underlying, impermeable bedrock. This replaces the unconfined aquifer parameterization (Niu et al. 2007) of previous versions with a zero flux lower boundary condition and an explicit water table depth (Lawrence et al. 2018). Sub-surface drainage is calculated as a function of an ice impedance factor, a baseflow calibration parameter, the topographic slope, and the thickness of the saturated part of the soil column (Lawrence et al. 2018). The distribution and temporal evolution of soil water within the soil column is calculated with a finite-difference approximation of the Richard's equation including Brooks-Corey parameterization. The hydraulic parameters involved in these calculations are determined by a weighted combination of mineral and organic properties. The mineral component of the soil hydraulic parameters is determined by pedotransfer functions and depends on sand and clay fractions (Clapp and Hornberger 1978). See appendix A for detailed equations of the pedotransfer function used in CLM5.

The numerical solution of the Richard's equation in CLM5 is based on a linearization that leads to a tridiagonal system of equations (Lawrence et al. 2018). CLM5 uses an adaptive time-stepping solver (Clark & Kavetski 2010, Kavetski et al. 2001) that improves the numerical stability for frozen soils and shallow bedrock compared to solvers in previous versions.

2.2 Data assimilation framework

2.2.1 Ensemble Kalman Filter

In Earth sciences, two common data assimilation approaches are variational methods, often used in atmospheric models, and sequential methods like the Ensemble Kalman filter (Reichle 2008). The Kalman Filter originates in filtering and prediction of linear dynamic systems (Kalman 1960) and the Ensemble Kalman Filter (EnKF) is a stochastic approximation for nonlinear dynamic systems based on Monte Carlo methods (Evensen 1994, Burgers et al. 1998). Included in PDAF are implementations of the most common variants of the Kalman filter. This study uses exclusively the ensemble Kalman filter (EnKF), in which an ensemble of independent model simulations is used to approximate the model error covariance matrix from the spread of the ensemble. For nonlinear models, like CLM5, ensemble spread is created from perturbations of model parameters and model forcings individually for each ensemble member. During the simulations, the EnKF uses an update step to assimilate observational data at time steps where observations are available. The update step is described by the following equation:

$$\mathbf{x}_a^i = \mathbf{x}_f^i + \mathbf{K}[\mathbf{y} - \mathbf{H}\mathbf{x}_f^i] \quad (1)$$

https://doi.org/10.5194/gmd-2021-38
Preprint. Discussion started: 8 September 2021

Geoscientific Model Development Discussions

where the superscript i refers to ensemble member i, $\mathbf{x}_a^i$ is the updated state vector after the analysis, $\mathbf{x}_f^i$ is the forecasted model state vector, **K** is the Kalman gain, **y** is the observation vector, and **H** is the so-called measurement operator that transforms between model and observational states. The Kalman gain **K** represents the weighting of observations versus model and is computed as follows:

$$\mathbf{K} = \mathbf{P}\mathbf{H}^{\mathrm{T}} \left(\mathbf{R} + \mathbf{H}\mathbf{P}\mathbf{H}^{\mathbf{T}}\right)^{-1} \quad (2)$$

where the superscript T refers to transposed matrices, **P** is the model error covariance matrix and **R** is the observational error covariance matrix. Therefore, the Kalman gain represents how much the model error contributes to the total error. Conceptually, **K** approaches 1 if the observational error covariance is very small compared to the model error covariance which in Eq. 1 would result in more weight for the correction based on the observational data. On the other hand, **K** approaches 0 if the observational error covariance is much larger than the model error covariance resulting in a smaller weight for the update term in Eq. 1. The observational error covariance matrix **R** is often statically defined based on the measurement error of the observations which are usually assumed to be independent. The model error covariance matrix **P** in the Ensemble Kalman Filter is approximated using the ensemble statistics. Specifically,

$$\mathbf{P} = \frac{1}{(\mathrm{N}-1)} \sum_{\mathrm{i}=1}^{\mathrm{N}} \left(\mathbf{x}_{\mathrm{f}}^{\mathrm{i}} - \overline{\mathbf{x}_{\mathrm{f}}}\right)\left(\mathbf{x}_{\mathrm{f}}^{\mathrm{i}} - \overline{\mathbf{x}_{\mathrm{f}}}\right)^{\mathrm{T}} \quad (3)$$

where N is the number of ensemble members and $\bar{x}$ is the ensemble mean. For example, ensemble members can be generated based on perturbed soil parameters and atmospheric forcings. The perturbations of soil properties and forcings represent the uncertainty range of the model.

Only during the data assimilation update step the ensemble members are connected through Eq. 3. Therefore, the ensemble Kalman filter is well-suited for parallelization. See Kurtz et al. (2016) for a discussion of the scaling of the Ensemble Kalman filter in PDAF.

Observational data is also perturbed for each ensemble member to maintain the correct error statistics (Burgers et al. 1998). Therefore, y in equation 1 is shorthand for y=o+i where o is the observational data and i is a perturbation vector with mean zero and covariance according to the observational error covariance matrix. Each ensemble member is independently propagated in time.

In this study, the observation vector y contains the soil water content observations, described in Section 3.2. The state vector $\mathbf{x}^i$ contains soil water content (model states), sand and clay fractions (parameters), and organic matter fractions (parameters) depending on the experiment as described in Section 3.3. The measurement operator in this case is a simple mapping of the three observation vector components to the state vector component at the corresponding depth.

2.2.2 Parameter updating

In this study, we also apply a joint state and parameter estimation approach to further improve simulation results. Specifically, the state augmentation approach (Friedland, 1969; Fertig et al., 2009) is applied in which the forecasted model state vector ($\mathbf{x}_f^i$

https://doi.org/10.5194/gmd-2021-38
Preprint. Discussion started: 8 September 2021


Geoscientific Model Development Discussions

in Eq. 1) contains both the model state variables and relevant model parameters. The attached model state parameters are updated based on the Kalman gain (Eq. 2) without direct observations of the model parameters.

For assimilation of soil water content the relevant model parameters are the hydraulic parameters. A common approach (Naz et al., 2019) is to indirectly update the hydraulic parameters by updating the soil texture, i.e. sand and clay fraction, and using the pedotransfer function as described in Section 2.1. Previous to CLM version 4.0 only sand and clay fractions were used to calculate the hydraulic parameters and therefore previous couplings of CLM and PDAF did not include organic matter as an option for joint state and parameter estimation. Similar to the work of Han et al. (2014) for CLM 4.5, we added organic matter as an additional parameter which can be updated with the CLM5+PDAF coupled model.

2.3 Coupling CLM5 with PDAF

As previously mentioned, this study makes use of the highly modular nature of TSMP (Shrestha et al. 2014) to integrate CLM5 as a new option for the land surface model component in the coupling framework. TSMP is designed to couple combinations of an atmospheric model, e.g. COSMO (Baldauf et al. 2011), a land surface model, e.g. CLM (Oleson et al. 2008), a sub-surface model, e.g. ParFlow (Ashby and Falgout 1996; Kollet and Maxwell 2006), and a data assimilation framework, e.g. PDAF (Nerger et al. 2005). The modularity allows not only the realization of a fully coupled system of all components, but also combinations like CLM and ParFlow or CLM and PDAF and also individual model components can be executed.

This study focuses on the implementation of the coupling of CLM5 and PDAF inside the TSMP framework. However, an advantage of implementing this single pair coupling inside a larger, modular platform is to facilitate future coupling implementations to the other components of TSMP. In general, the coupling in TSMP uses the Ocean-Atmosphere-Sea-Ice-Soil coupler – Model Coupling Toolkit (OASIS-MCT) (Valcke et al., 2013) to couple the models in a multiple program multiple data (MPMD) approach. However, as described in Kurtz et al. (2016), coupling with PDAF is an exception to this approach. Instead of using MPMD, a single executable is built out of modified, pseudo-library versions of the models. This keeps all model data in main memory and avoids I/O intensive re-initialization of models. Additionally, since in this study only one model (CLM5) and PDAF are coupled, the utilization of OASIS-MCT is not necessary.

Figure 1 shows the five main components necessary for coupling CLM5 and PDAF in the TSMP framework and their connection. The PDAF components, core functions and user functions, are the same as described in Nerger et al. (2005) and Kurtz et al. (2016) respectively. The only modifications to code in the PDAF user functions are superficial inclusions of CLM5 as option with the same functionality as already implemented and described by Kurtz et al. (2016) for CLM 3.5.

The main program, labeled TSMP-PDAF driver, controls the individual components and handles the parallel communication using multiple MPI communicators. Adding CLM5 coupling requires only minor changes to the TSMP-PDAF driver to add CLM5 as a new option to the models controlled by the driver.

https://doi.org/10.5194/gmd-2021-38
Preprint. Discussion started: 8 September 2021


Geoscientific Model Development Discussions

The TSMP wrapper contains the majority of additional code for coupling CLM5 and PDAF. The TSMP-PDAF driver uses the TSMP wrapper as an interface to the individual pseudo-libraries of the models. Therefore, the TSMP wrapper contains the modified routines from the model for initialization, time stepping, and clean-up. These routines are moved from the CLM5 specific CIME driver into the TSMP wrapper. The clean-up routine is migrated without modification. The modification to the initialization routine involves an added call to the subroutine that defines the state vector. The main time stepping loop in CLM5 works by looping until a stop alarm is received. On the other hand, the TSMP framework, similar to older versions of CLM, works with a loop counting up until a specified end time is reached enabling data assimilation at specified time steps. Therefore, the TSMP wrapper subroutine to advance CLM5 contains only the code from inside the original time stepping loop. In this way, the TSMP-PDAF driver can control how many CLM5 time steps are performed before stopping for an interrupting data assimilation step. Further modifications to the time stepping subroutine include the addition of calling the PDAF specific subroutine to set the state vector before each data assimilation step.

Additionally, the TSMP wrapper contains the model specific routines for managing the PDAF state vector. This includes defining the size of the state vector based on domain decomposition, for non-single grid cell simulations and options for parameter updating. The TSMP wrapper provides both the subroutine called by the model to set the state vector and the subroutine called by the data assimilation method to update the model variables contained in the state vector. For soil water content and soil texture parameters setting the state vector is simply copying the model values to their respective place in the state vector. The subroutine to update the state vector contains functionality to catch invalid values, e.g. below residual soil water content, above porosity, and below 0% or above 100% for the sum of the sand and clay fractions. Furthermore, for the optional parameter updating it is necessary to provide a function to transform the input parameters, e.g. soil texture, to the model parameters, e.g. the soil hydraulic parameters. CLM5 performs this transformation once during initialization to obtain the hydraulic parameters from the soil texture in the surface file. As mentioned in Section 2.1, this procedure has changed compared to older versions of CLM. The subroutine to perform this transform after each data assimilation step follows the implementation in CLM5 and is shown in Appendix A.

The component labeled libclm5 in Figure 1 is the pseudo-library from CLM5 compiled modules. Code modifications for CLM5 source files are limited to two driver modules related to parallel communication and ensemble reading of namelist files. As previously mentioned, the TSMP-PDAF driver manages the initialization of the parallel communication that involves initializing MPI and splitting the global communicator MPI_COMM_WORLD into specific model, filter, and coupling communicators. However, by default CLM5 also initializes MPI and uses MPI_COMM_WORLD for its parallel communication. Since only one MPI_COMM_WORLD can exist within a MPI application, the CLM5 code was modified to not initialize MPI and not use MPI_COMM_WORLD.

For ensemble simulations, each ensemble member has individual input files. In CLM input files are controlled by namelists. In older versions of CLM a single namelist was used and to enable ensemble simulations for TSMP-PDAF only involved attaching an ensemble identifier suffix to the name of this namelist. In CLM5 there are multiple namelists and managing the

https://doi.org/10.5194/gmd-2021-38
Preprint. Discussion started: 8 September 2021


Geoscientific Model Development Discussions

reading of them has become more complex. However, CLM5 also supports an ensemble mode where each ensemble member reads namelists with identifier suffixes. Our implementation of CLM5+PDAF makes use of this ensemble mode. The ensemble mode is modified such that it uses the PDAF model communicator instead of splitting the global communicator. Therefore, the initialization subroutine that handles the ensemble mode is modified to accept a communicator and an individual ensemble member number from PDAF. Additionally, the initialization subroutine also passes the PDAF information to the subroutine that initializes the communicators for CLM5 and replaces the default ensemble mode identifiers with the PDAF specific identifiers. Figure 2 illustrates these modifications and shows the general process flow difference between CLM5 and CLM5+PDAF, i.e., the interruption of the CLM simulation by the PDAF data assimilation step.

3. Test case

3.1 Study Site

The coupled modeling framework is applied to the small (38.5 ha) forested catchment Wüstebach located in the Eifel National Park near the German-Belgian border. As part of the Terrestrial Environmental Observatories (TERENO) network (Bogena et al., 2015; Bogena et al., 2018), the Wüstebach site uses a wireless sensor network (SoilNet) to provide soil water content and soil temperature measurements since 2009 at 5cm, 20cm, and 50cm depth at 150 locations every 15 minutes (Bogena et al. 2010).

The Wüstebach test site is also interesting because in the late summer / early autumn of 2013 the national park forest management removed the prevailing spruce monoculture forest in an area to promote the natural regeneration of deciduous forest. The SoilNet was installed before this change, so that the impact on the soil water content is measured before and after this land-use change. However, in this study we use the study site mainly to demonstrate the functionality of the newly coupled CLM5+PDAF framework and therefore, we focus on the undisturbed forested area.

As mentioned in the introduction, we do not focus on spatial heterogeneity but instead look at the study site as it would be modeled in a regional or continental simulation, i.e., as a single grid cell. This allows for a clear and simple setup to test and demonstrate the functionality of CLM5+PDAF and simultaneously allows us to use a larger ensemble than is usually feasible for regional or continental data assimilation simulations. More specific details on the simulation setup are presented in Section 3.3.

https://doi.org/10.5194/gmd-2021-38
Preprint. Discussion started: 8 September 2021


Geoscientific Model Development Discussions

3.2 Data

3.2.1 Soil water content – in-situ measurements

The observational data of the study site Wüstebach is pre-processed before assimilation. The raw data from the TERENO data portal (Sorg et al. 2015) contains data for all stations and all sensors in 15 minutes intervals including quality flags. The observational data is pre-processed using filters that remove data based on their quality flag, spikes, frozen soil condition, and erroneous values. Spikes are defined as reductions in soil water content of more than 1 vol% or increases in soil water content of more than 5 vol% with an immediate return to values within 1% of the value before the spike. Soil water content below 1 vol.% or above 90 vol.% is considered erroneous. These thresholds and the definition of spikes are based on Wiekenkamp et al. (2016) and Dorigo et al. (2013). In Wüstebach each soil water content sensor is paired with a soil temperature sensor. This allows for the removal of unreliable measurements due to frozen soil. Time steps in which less than 25% of all sensors provide data are filtered out. The filtered raw data is then spatially and temporally averaged to fit the requirements of the model, i.e., daily averages for the three soil depths.

As mentioned above, the Wüstebach was partially deforested in 2013, with SoilNet SWC sensors covering both the undisturbed and deforested areas. The deforested part of the Wüstebach catchment is mainly located in the riparian zone featuring shallow groundwater that is strongly influenced by incoming lateral flows within the catchment. However, lateral flows are not well represented in the single point CLM5 setup. Therefore, we omitted the riparian zone and selected only SoilNet stations located in the groundwater distant forested parts of the Wüstebach catchment in this study. With these criteria 37 soil water stations remain in the forested part of the Wüstebach catchment and are used in this study.

3.2.2 Atmospheric forcings

The atmospheric forcings used in this study are measurements of air pressure, shortwave radiation, relative humidity, 2m air temperature, and wind speed from an on-site meteorological station. Additionally, the precipitation data is provided by the meteorological station Kaltenherberg (DWD, German Weather Service) located 5km west of the Wüstebach study site (Bogena et al., 2015). The atmospheric forcing data is perturbed to generate an ensemble for data assimilation using the EnKF. In this study, the perturbed variables are precipitation, shortwave radiation, longwave radiation, and air temperature. These variables are perturbed according to cross-correlation coefficients derived from global observations by Reichle et al. (2007). The specific perturbation characteristics used in this study are from Han et al. (2014) and shown in Table 1.

3.2.3 Surface parameters

The over 70 different surface parameters included in each CLM5 surface file are generated by the tools provided by CLM5 from remapping of various pre-processed global files, see Lawrence et al. (2019) for details. For the single grid cell of the

https://doi.org/10.5194/gmd-2021-38
Preprint. Discussion started: 8 September 2021


Geoscientific Model Development Discussions

study site, all default values were used, except for the plant functional type and the depth to bedrock. We chose the plant functional type "needleleaf evergreen temperate tree" to represent the spruce monoculture of the Wüstebach site. The depth to bedrock was adjusted to 1.6 meters according to Fang et al. (2015). Sand, clay, and organic matter fractions are perturbed for each ensemble member. Perturbed values were obtained by drawing from a uniform distribution with mean zero and a range between -20% and +20%. Perturbations that cause the sum of sand and clay fractions to exceed 100% are re-scaled to be limited to 100%. These perturbations are larger than, for example, the ones used in Han et al. (2014) to represent a larger initial model parameter uncertainty for a single grid cell simulation with a larger ensemble.

3.3 Simulation experiments

Four different setups were used to demonstrate the functionality and effectiveness of CLM5+PDAF. The open loop (OL) setup has forward simulations without data assimilation. These simulations are equivalent to CLM5 standalone ensemble simulations with perturbed inputs. The initial data assimilation setup limits the state vector to the soil water content variable (DA_s). The data assimilation with state and parameter updates setup (DA_s+p) applies the joint state and parameter estimation approach, described in Section 2.2.2, by augmenting the state vector with sand and clay fractions. The fourth setup, data assimilation with state and parameter updates including organic matter (DA_s+p+o) adds the soil organic matter fraction to the state vector. All setups were run for a 10 year time period starting from 2009 when observations become available.

We used four statistical metrics to evaluate the quality of the simulation results: the root-mean-square-error (RMSE), the unbiased root-mean-square-error (ubRMSE), the mean bias error (MBE) and the squared correlation coefficient (R^2):

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(\mathbf{Hx}^i - \mathbf{y}^i)^2}{N}} \quad (4)$$

$$\mathrm{ubRMSE} = \sqrt{\frac{\sum_{i=1}^{N}[(\mathbf{Hx}^i - \overline{\mathbf{Hx}^i}) - (\mathbf{y}^i - \overline{\mathbf{y}^i})]^2}{N}} \quad (5)$$

$$\mathrm{MBE} = \frac{\sum_{i=1}^{N}(\mathbf{Hx}^i - \mathbf{y}^i)}{N} \quad (6)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(\mathbf{y}^i - \mathbf{Hx}^i)^2}{\sum_{i=1}^{N}(\mathbf{y}^i - \overline{\mathbf{y}^i})^2} \quad (7)$$

where **y** represents observations, **Hx** represents simulated values, i is the ensemble member, N the total number of ensemble members and overbar represents ensemble average.

https://doi.org/10.5194/gmd-2021-38
Preprint. Discussion started: 8 September 2021


Geoscientific Model Development Discussions

3.4 Comparison of the four different simulation setups

Figure 3 shows time series of the monthly averaged SWC at the three observation depths. The monthly averages highlight better the tendencies for the different simulation setups. The simulation results of all model setups show a good agreement with the soil water content observations at 20 cm depth, while there are clear deviations at 5 cm and 50 cm depth. The simulations tend to overestimate SWC compared to the observations. This difference is more distinct in the summer months at 5 cm depth, especially during the dry summer of 2018. The soil water content overestimation by the model at 50 cm depth is smaller, but more consistent over time. The results of the simulations at 5 cm depth illustrate the effects of data assimilation: All three data assimilation setups provide soil water content predictions that are closer to the observations compared to the open loop setup simulation.

The scatter plots in Figures 4, 5, and 6 show the comparisons of the different data assimilation scenarios in terms of correlation of daily soil water content averages between observations and simulations. Table 2 summarizes the complementary statistical results. The evaluation at 5 cm depth, shown in the top left of Figure 4, reflects the overestimation of soil moisture content by the open loop simulation. All observed daily average SWC below 40% are overestimated by the model. The other three scatter plots in Figure 4 highlight the progressive effectiveness of the three data assimilation setups. While the DA_s setup still shows overestimation of SWC compared to observations it reduces the RMSE compared to the OL setup by 30%, the ubRMSE by 35%, and increases the R^2 to above 0.9. The DA_s+p and DA_s+p+o setups show similar, improved results. DA_s+p performs slightly better in terms of ubRMSE and R^2 than the DA_s+p+o but slightly worse in terms of RMSE and MBE.

The results at 20 cm depth (Fig. 5) show a closer agreement between observations and simulations than the results at 5cm and 50cm depth. At 20cm depth, simulations slightly underestimate SWC. Similar to 5cm depth, the DA_s improves the RMSE by 30% compared to OL and increases the R^2 to above 0.9. At 20 cm depth, the DA_s+p+o shows an especially small MBE and overall very good agreement with the observations, suggesting that updating the organic matter faction does contribute to more accurate simulation results.

The results from 50 cm depth (Fig. 6) show the most consistent overestimation of SWC by the model and the smallest improvement by data assimilation. The DA_ssetup reduces the RMSE by only 7% compared to the OL and even the best performing setup (DA_s+p+o) only improves RMSE by 15%. The DA_s+p and DA_s+p+o scenarios result in similar results at 5cm depth (Fig. 4).

Table 3 shows the changes in soil texture related to the parameter updates. The parameter updates increase the sand fraction at all three measurement depths by a factor of 2. The clay fraction, on the other hand, is only slightly reduced across these depths. Organic matter fraction is also increased in all three depths, but more significantly in 5 cm and 20cm.

https://doi.org/10.5194/gmd-2021-38
Preprint. Discussion started: 8 September 2021


Geoscientific Model Development Discussions

4. Conclusions

In this study, we presented the newly coupled data assimilation framework CLM5+PDAF. The presented implementation can be summarized by three main aspects: the online variant of PDAF, re-use of CLM5 ensemble mode, and the TSMP framework. The PDAF online variant performs data assimilation in the main memory during runtime by coupling the model and PDAF in a single executable. We have described the necessary code modifications to achieve this coupling. The presented implementation re-uses the CLM5 ensemble mode which enables multiple simulations to run in parallel from the same executable while using independent inputs and creating individual outputs. This re-use minimizes necessary code changes to connect CLM5 and PDAF and simplifies the management of the parallel communicators of CLM5 and PDAF. The framework of TSMP provided the build infrastructure and the template for the coupling components. We chose to include CLM5+PDAF in the TSMP to make it available for future developments in the modular environment and facilitate future couplings to other components. The performance of the CLM5+PDAF data assimilation system was illustrated with the assimilation of soil water content data for the Wüstebach site in Germany. Data assimilation decreases the mismatch between observations and model states. We further showed that including parameter updates can improve overall estimations, although some systematic bias remains. Updating also organic matter fraction, as one of the parameters determining the soil hydraulic properties, has an overall positive effect. However, even with this addition some significant differences between simulated and observed values remain, especially at 5cm depth and in dry years.

The performance of CLM5+PDAF can be further improved in the future by updating soil hydraulic parameters themselves, instead of indirectly updating them via soil texture and pedotransfer functions. This could potentially reduce the model uncertainty further since the accuracy of the pedotransfer functions would be less of an issue after parameter updating. This will require more fundamental code changes and will be considered in future work. In addition, CLM5+PDAF will be further extended by the assimilation of more state variables, like for example LAI or soil temperature.

Appendix A: CLM5 specific equations relating sand, clay, and organic matter fractions to soil hydraulic parameters

In CLM5 the soil hydraulic parameters are determined by a weighted average of the respective mineral and organic components. Specifically, for the mineral component the following approximations from Cosby et al. (1984) are used:

$$\theta_{(\mathrm{min,sat,i})} = 0.489 - 0.00126\ (\%\mathrm{sand})_i \quad \text{(A1)}$$

where $\theta_{(\mathrm{min,sat,i})}$ is the porosity of the mineral part and subscript i refers to the vertical level.

$$B_{(\mathrm{min,i})} = 2.91 + 0.159\ (\%\mathrm{clay})_i \quad \text{(A2)}$$

where $B_{(\mathrm{min,i})}$ is the hydraulic conductivity exponent of the mineral part.

$$k_{(\mathrm{min,sat,i})} = 0.0070556\ (10^{\,-0.884\,+\,0.0153\ (\%\mathrm{sand})_i}) \quad \text{(A3)}$$

https://doi.org/10.5194/gmd-2021-38
Preprint. Discussion started: 8 September 2021


Geoscientific Model Development Discussions

where $k_{(min,sat,i)}$ is the saturated hydraulic conductivity of the mineral part.

$$\Psi_{(min,sat,i)} = 10\ (10^{(1.88-0.0131\ (\%sand)_i}\ \text{(A4)}$$

where $\Psi_{(min,sat,i)}$ is the saturated suction / saturated soil matric potential of the mineral part and is related to the adsorptive and capillary forces within the soil matrix.

The organic component of the soil hydraulic parameters is approximated by the following equations from Lawrence and Slater (2008):

$$\theta_{(om,sat,i)} = \max(0.83, 0.93 - 0.1\ D_i)\ \text{(A5)}$$

Where $\theta_{(om,sat,i)}$ is the porosity of the organic part and

$$D_i = \frac{depth_i}{zsapric}\ \text{(A6)}$$

where $depth_i$ is the depth of the vertical level and zsapric is the depth at which organic matter takes on characteristics of sapric peat.

$$B_{(om,i)} = \max(12, 2.7 + 9.3\ D_i)\ \text{(A7)}$$

where$B_{(om,i)}$ is the hydraulic conductivity exponent for the organic part.

$$k_{(om,sat,i)} = \max(k_{(min,sat,i)}, 0.28 - 0.2799\ D_i)\ \text{(A8)}$$

where $k_{(om,sat,i)}$ is the saturated hydraulic conductivity for the organic part.

$$\Psi_{(om,sat,i)} = \min\ (10.1, 10.3 - 0.2\ D_i)\ \text{(A9)}$$

where $\Psi_{(om,sat,i)}$ is the saturated suction of the organic part.

Code availability.

The development branch of the CLM5+PDAF coupling is freely available via Zenodo, doi:10.5281/zenodo.4534157

Data availability.

Soil water content data from the TERENO site Wüstebach (TERENO ID: WU_B_001 to WU_B_150) are freely available via the TERENO data portal TEODOOR (http://teodoor.icg.kfa-juelich.de/).

https://doi.org/10.5194/gmd-2021-38
Preprint. Discussion started: 8 September 2021


Geoscientific Model Development Discussions

Author contribution.

L.S. pre-processed the data, developed the code, designed and performed the simulations, and prepared the manuscript. H.B., H.J.H.F. and H.V. supervised the research, co-designed the experiments, and contributed to the manuscript.

Competing interests.

The authors declare that they have no conflict of interest.

Acknowledgements.

The authors gratefully acknowledge the support by the project LIFE RESILIENT FORESTS – Coupling water, fire and climate resilience with biomass production from forestry to adapt watersheds to climate change. This project is co-funded by the LIFE Programme of the European Union under contract number LIFE 17 CCA/ES/000063. Furthermore, the authors gratefully acknowledge the computing time granted through JARA on the supercomputer JURECA at Forschungszentrum Jülich. This work used data provided by the Helmholtz Association and the Federal Ministry of Education and Research (BMBF) in the framework of TERENO (Terrestrial Environmental Observatories).

References

Anderson, J., Hoar, T., Raeder, K., Liu, H., Collins, N., Torn, R., and Avellano, A.: The data assimilation research testbed: A community facility, Bulletin of the American Meteorological Society, 90, 1283–1296, 2009.

Ashby, S. F. and Falgout, R. D.: A parallel multigrid preconditioned conjugate gradient algorithm for groundwater flow simulations, Nuclear science and engineering, 124, 145–159, 1996.

Baldauf, M., Seifert, A., Förstner, J., Majewski, D., Raschendorfer, M., and Reinhardt, T.: Operational convective-scale numerical weather prediction with the COSMO model: Description and sensitivities, Monthly Weather Review, 139, 3887–3905, 2011.

Beven, K. J. and Kirkby, M. J.: A physically based, variable contributing area model of basin hydrology/Un modèle à base physique de zone d'appel variable de l'hydrologie du bassin versant, Hydrological Sciences Journal, 24, 43–69, 1979.

Bogena, H., Herbst, M., Huisman, J., Rosenbaum, U., Weuthen, A., and Vereecken, H.: Potential of wireless sensor networks for measuring soil water content variability, Vadose Zone Journal, 9, 1002–1013, 2010.

Bogena, H., Bol, R., Borchard, N., Brüggemann, N., Diekkrüger, B., Drüe, C., Groh, J., Gottselig, N., Huisman, J., Lücke, A., et al.: A terrestrial observatory approach to the integrated investigation of the effects of deforestation on water, energy, and matter fluxes, Science China Earth Sciences, 58, 61–75, 2015.

https://doi.org/10.5194/gmd-2021-38
Preprint. Discussion started: 8 September 2021

Bogena, H., Montzka, C., Huisman, J., Graf, A., Schmidt, M., Stockinger, M., Von Hebel, C., Hendricks-Franssen, H., Van der Kruk, J., Tappe, W., et al.: The TERENO-Rur hydrological observatory: A multiscale multi-compartment research platform for the advancement of hydrological science, Vadose Zone Journal, 17, 1–22, 2018.

Buotte, P. C., Levis, S., Law, B. E., Hudiburg, T. W., Rupp, D. E., and Kent, J. J.: Near-future forest vulnerability to drought and fire varies across the western United States, Global change biology, 25, 290–303, 2019.

Burgers, G., Jan van Leeuwen, P., and Evensen, G.: Analysis scheme in the ensemble Kalman filter, Monthly weather review, 126, 1719–1724, 1998. Clapp, R. B. and Hornberger, G. M.: Empirical equations for some soil hydraulic properties, Water resources research, 14, 601–604, 1978.

Clark, M. P. and Kavetski, D.: Ancient numerical daemons of conceptual hydrological modeling: 1. Fidelity and efficiency of time stepping schemes, Water Resources Research, 46, 2010.

Cornelissen, T., Diekkrüger, B., and Bogena, H. R.: Significance of scale and lower boundary condition in the 3D simulation of hydrological processes and soil moisture variability in a forested headwater catchment, Journal of hydrology, 516, 140–153, 2014.

Cornelissen, T., Diekkrüger, B., and Bogena, H. R.: Using high-resolution data to test parameter sensitivity of the distributed hydrological model HydroGeoSphere, Water, 8, 202, 2016.

Cosby, B., Hornberger, G., Clapp, R., and Ginn, T.: A statistical exploration of the relationships of soil moisture characteristics to the physical properties of soils, Water resources research, 20, 682–690, 1984.

Dorigo, W., Xaver, A., Vreugdenhil, M., Gruber, A., Hegyiova, A., Sanchis-Dufau, A., Zamojski, D., Cordes, C., Wagner, W., and Drusch, M.: Global automated quality control of in situ soil moisture data from the International Soil Moisture Network, Vadose Zone Journal, 12, 2013.

Evensen, G.: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics, Journal of Geophysical Research: Oceans, 99, 10 143–10 162, 1994.

Fang, Z., Bogena, H., Kollet, S., Koch, J., and Vereecken, H.: Spatio-temporal validation of long-term 3D hydrological simulations of a forested catchment using empirical orthogonal functions and wavelet coherence analysis, Journal of hydrology, 529, 1754–1767, 2015.

Fang, Z., Bogena, H., Kollet, S., and Vereecken, H.: Scale dependent parameterization of soil hydraulic conductivity in 3D simulation of hydrological processes in a forested headwater catchment, Journal of hydrology, 536, 365–375, 2016.

Fertig, E., Baek, S.-J., Hunt, B., Ott, E., Szunyogh, I., Aravéquia, J., Kalnay, E., Li, H., and Liu, J.: Observation bias correction with an ensemble Kalman filter, Tellus A: Dynamic Meteorology and Oceanography, 61, 210–226, 2009.

Fox, A. M., Hoar, T. J., Anderson, J. L., Arellano, A. F., Smith, W. K., Litvak, M. E., MacBean, N., Schimel, D. S., and Moore, D. J.: Evaluation of a data assimilation system for land surface models using CLM4. 5, Journal of Advances in Modeling Earth Systems, 10, 2471–2494, 2018.

Friedland, B.: Treatment of bias in recursive filtering, IEEE Transactions on Automatic Control, 14, 359–367, 1969.

https://doi.org/10.5194/gmd-2021-38
Preprint. Discussion started: 8 September 2021

Han, X., Franssen, H.-J. H., Montzka, C., and Vereecken, H.: Soil moisture and soil properties estimation in the Community Land Model with synthetic brightness temperature observations, Water resources research, 50, 6081–6105, 2014.

Hudiburg, T., Law, B., and Thornton, P.: Evaluation and improvement of the Community Land Model (CLM4) in Oregon forests, Biogeosciences, 10, 453–470, 2013.

Kalman, R. E.: A new approach to linear filtering and prediction problems, 1960.

Kavetski, D., Binning, P., and Sloan, S.: Adaptive time stepping and error control in a mass conservative numerical solution of the mixed form of Richards equation, Advances in water resources, 24, 595–605, 2001.

Koch, J., Cornelissen, T., Fang, Z., Bogena, H., Diekkrüger, B., Kollet, S., and Stisen, S.: Inter-comparison of three distributed hydrological models with respect to seasonal variability of soil moisture patterns at a small forested catchment, Journal of hydrology, 533, 234–249, 2016.

Kollet, S. J. and Maxwell, R. M.: 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.

Kurtz, W., He, G., Kollet, S. J., Maxwell, R. M., Vereecken, H., and Hendricks Franssen, H.-J.: 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.

Lawrence, D., Fisher, R., Koven, C., Oleson, K., Swenson, S., Vertenstein, M., Andre, B., Bonan, G., Ghimire, B., van Kampenhout, L., et al.: Technical description of version 5.0 of the Community Land Model (CLM), National Center for Atmospheric Research, University Corporation for Atmospheric Research, Boulder, CO, 2018.

Lawrence, D. M. and Slater, A. G.: Incorporating organic soil into a global climate model, Climate Dynamics, 30, 145–160, 2008.

Lawrence, D. M., Fisher, R. A., Koven, C. D., Oleson, K. W., Swenson, S. C., Bonan, G., Collier, N., Ghimire, B., van Kampenhout, L., Kennedy, D., et al.: The Community Land Model version 5: Description of new features, benchmarking, and impact of forcing uncertainty, Journal of Advances in Modeling Earth Systems, 11, 4245–4287, 2019.

Ling, X., Fu, C., Guo, W., and Yang, Z.-L.: Assimilation of remotely sensed LAI into CLM4CN using DART, Journal of Advances in Modeling Earth Systems, 11, 2768–2786, 2019.

Naz, B. S., Kurtz, W., Montzka, C., Sharples, W., Goergen, K., Keune, J., Gao, H., Springer, A., Hendricks Franssen, H.-J., and Kollet, S.: Improving soil moisture and runoff simulations at 3 km over Europe using land surface data assimilation, Hydrology and earth system sciences, 23, 277–301, 2019.

Naz, B. S., Kollet, S., Franssen, H.-J. H., Montzka, C., and Kurtz, W.: A 3 km spatially and temporally consistent European daily soil moisture reanalysis from 2000 to 2015, Scientific data, 7, 1–14, 2020.

Nerger, L., Hiller, W., and Schröter, J.: PDAF-the parallel data assimilation framework: experiences with Kalman filtering, in: Use of high performance computing in meteorology, pp. 63–83, World Scientific, 2005.

Niu, G.-Y., Yang, Z.-L., Dickinson, R. E., and Gulden, L. E.: A simple TOPMODEL-based runoff parameterization (SIMTOP) for use in global climate models, Journal of Geophysical Research: Atmospheres, 110, 2005.

https://doi.org/10.5194/gmd-2021-38
Preprint. Discussion started: 8 September 2021

Niu, G.-Y., Yang, Z.-L., Dickinson, R. E., Gulden, L. E., and Su, H.: Development of a simple groundwater model for use in climate models and evaluation with Gravity Recovery and Climate Experiment data, Journal of Geophysical Research: Atmospheres, 112, 2007.

Oleson, K., Niu, G.-Y., Yang, Z.-L., Lawrence, D., Thornton, P., Lawrence, P., Stöckli, R., Dickinson, R., Bonan, G., Levis, S., et al.: Improvements to the Community Land Model and their impact on the hydrological cycle, Journal of Geophysical Research: Biogeosciences, 113, 2008.

Overgaard, J., Rosbjerg, D., and Butts, M.: Land-surface modelling in hydrological perspective–a review, Biogeosciences, 3, 229–241, 2006.

Post, H., Vrugt, J. A., Fox, A., Vereecken, H., and Hendricks Franssen, H.-J.: Estimation of Community Land Model parameters for an improved assessment of net carbon fluxes at European sites, Journal of Geophysical Research: Biogeosciences, 122, 661–689, 2017.

Reichle, R. H., Koster, R. D., Liu, P., Mahanama, S. P., Njoku, E. G., and Owe, M.: Comparison and assimilation of global soil moisture retrievals from the Advanced Microwave Scanning Radiometer for the Earth Observing System (AMSR-E) and the Scanning Multichannel Microwave Radiometer (SMMR), Journal of Geophysical Research: Atmospheres, 112, 2007.

Reichle, R. H., Crow, W. T., and Keppenne, C. L.: An adaptive ensemble Kalman filter for soil moisture data assimilation, Water resources research, 44, 2008.

Shrestha, P., Sulis, M., Masbou, M., Kollet, S., and Simmer, C.: A scale-consistent terrestrial systems modeling platform based on COSMO, CLM, and ParFlow, Monthly weather review, 142, 3466–3483, 2014.

Sorg, J. and Kunkel, R.: Conception and implementation of an ogc-compliant sensor observation service for a standardized access to raster data, ISPRS International Journal of Geo-Information, 4, 1076–1096, 2015.

Valcke, S.: The OASIS3 coupler: A European climate modelling community software, Geoscientific Model Development, 6, 373–388, 2013.

Vereecken, H., Huisman, J., Bogena, H., Vanderborght, J., Vrugt, J., and Hopmans, J.: On the value of soil moisture measurements in vadose zone hydrology: A review, Water resources research, 44, 2008.

Wiekenkamp, I., Huisman, J. A., Bogena, H. R., Graf, A., Lin, H., Drüe, C., and Vereecken, H.: Changes in measured spatiotemporal patterns of hydrological response after partial deforestation in a headwater catchment, Journal of hydrology, 542, 648–661, 2016.

Wood, E. F., Roundy, J. K., Troy, T. J., Van Beek, L., Bierkens, M. F., Blyth, E., de Roo, A., Döll, P., Ek, M., Famiglietti, J., et al.: Hyperresolution global land surface modeling: Meeting a grand challenge for monitoring Earth's terrestrial water, Water Resources Research, 47, 2011.

Zhang, L., Lei, H., Shen, H., Cong, Z., Yang, D., and Liu, T.: Evaluating the representation of vegetation phenology in the Community Land Model 4.5 in a temperate grassland, Journal of Geophysical Research: Biogeosciences, 124, 187–210, 2019.

https://doi.org/10.5194/gmd-2021-38
Preprint. Discussion started: 8 September 2021

Geoscientific Model Development Discussions

Zhang, Y.-F., Hoar, T. J., Yang, Z.-L., Anderson, J. L., Toure, A. M., and Rodell, M.: Assimilation of MODIS snow cover through the Data Assimilation Research Testbed and the Community Land Model version 4, Journal of Geophysical Research: Atmospheres, 119, 7091–7103, 2014.

Zhao, L., Yang, Z.-L., and Hoar, T. J.: Global soil moisture estimation by assimilating AMSR-E brightness temperatures in a coupled CLM4–RTM–DART system, Journal of Hydrometeorology, 17, 2431–2454, 2016.

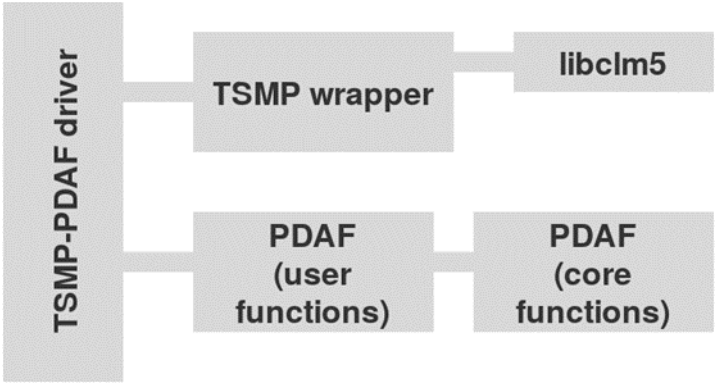


Figure 1: Components of TSMP CLM5+PDAF highlighting the distinct separation of PDAF functionality, TSMP wrapper, and CLM5 pseudo-library.

https://doi.org/10.5194/gmd-2021-38
Preprint. Discussion started: 8 September 2021

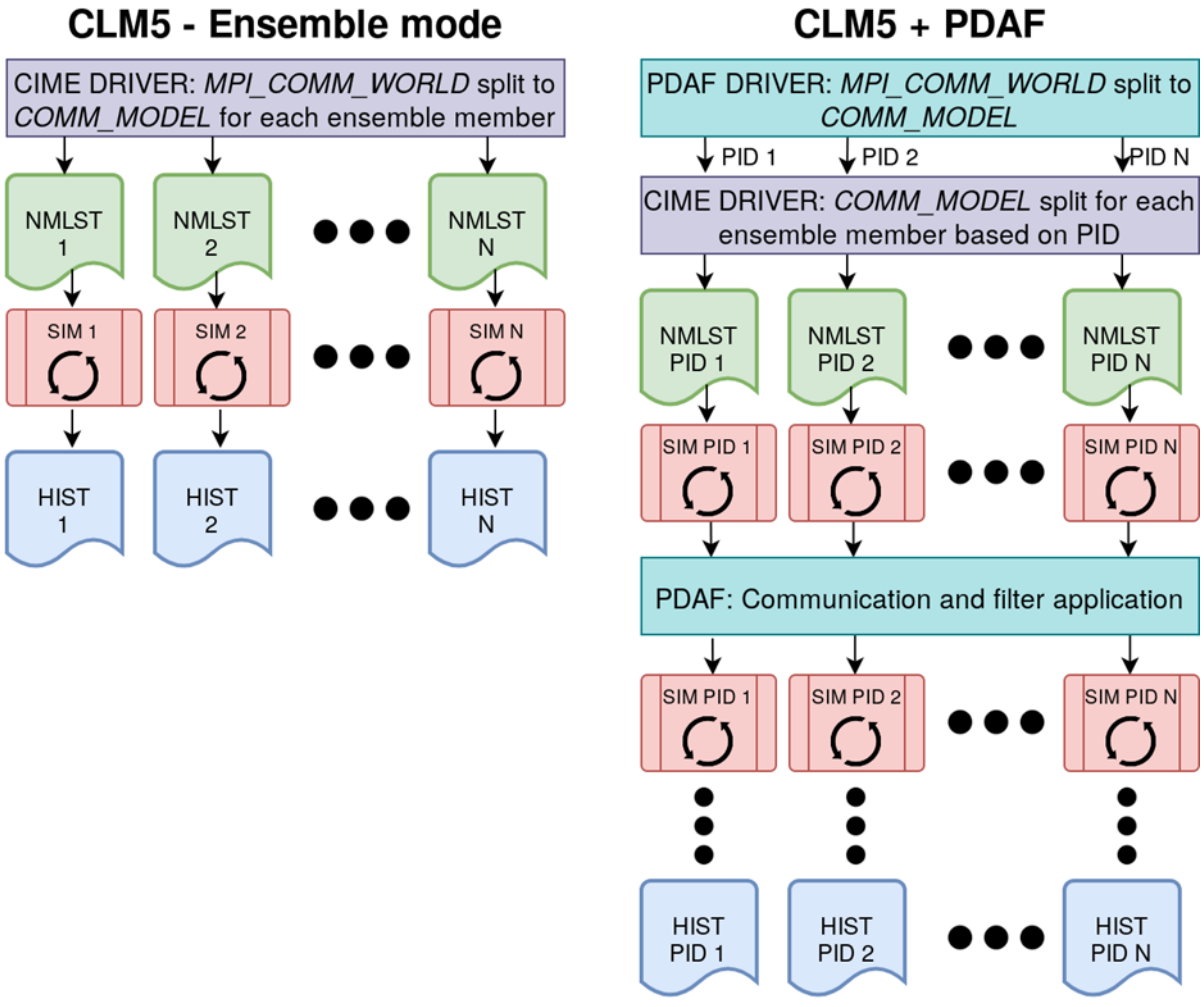


Figure 2: Schematic overview of CLM5 ensemble mode (left side) and CLM5+PDAF (right side) communication initialization and process flow. In the diagram NMLST means namelist, SIM means simulation process, HIST means history file output, PID means PDAF identification number.

https://doi.org/10.5194/gmd-2021-38
Preprint. Discussion started: 8 September 2021

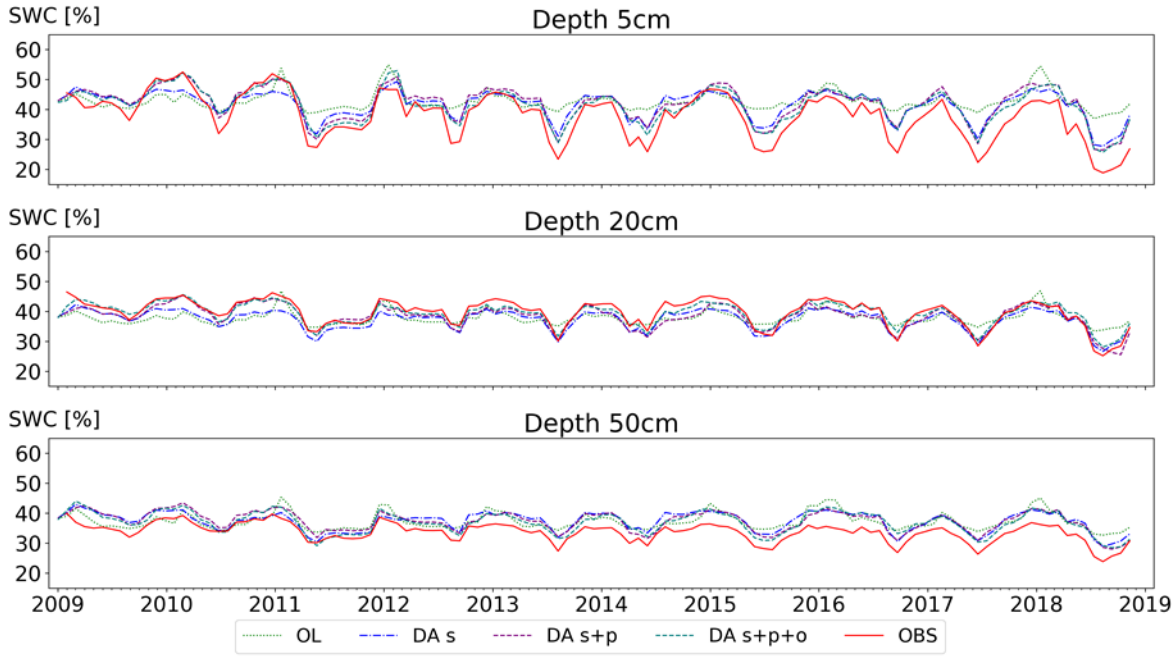


Figure 3: Time series of the monthly averaged soil water content (SWC) from 2009 to 2018 at the three different depths and for each simulation scenario. The red, full line shows observational data. The light green, dotted line shows open loop simulation results. The blue, dash-dotted line shows results for data assimilation of state variables. The purple, dashed line shows results for the assimilation of states and updating of parameters. The dark green, dashed line shows results for assimilation of states and updating of parameters including organic matter.

https://doi.org/10.5194/gmd-2021-38
Preprint. Discussion started: 8 September 2021

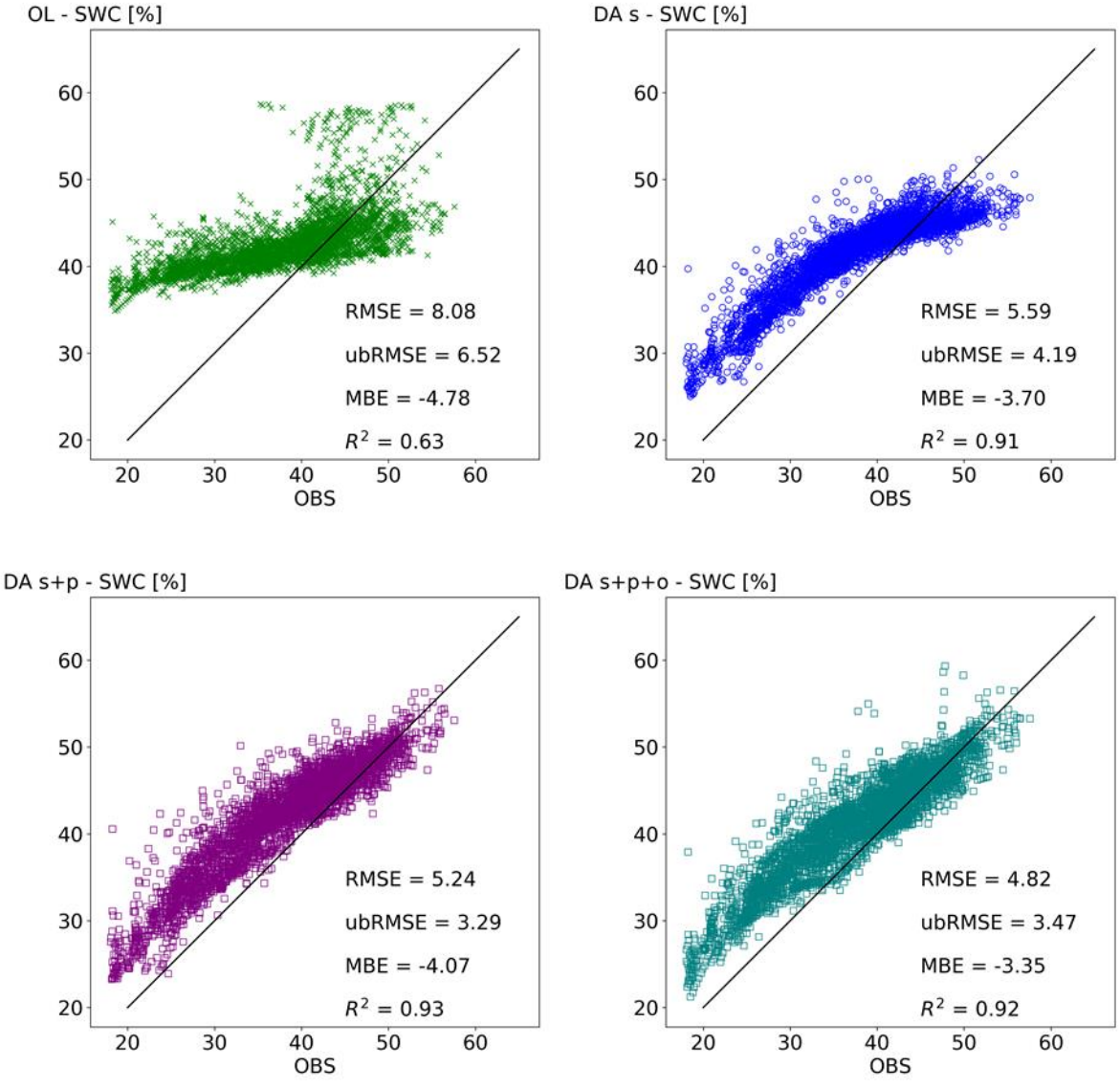
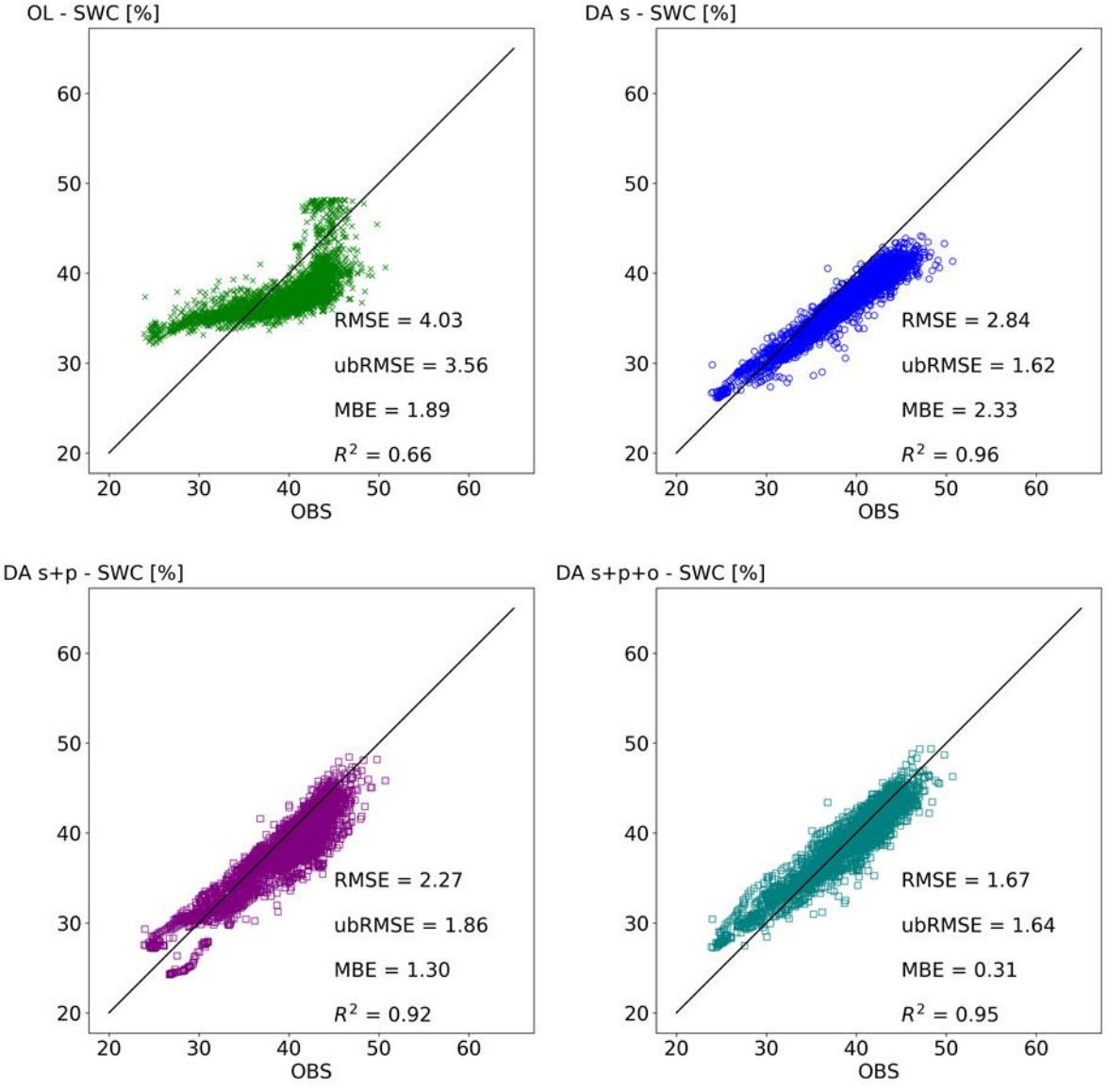


Figure 4: Correlation diagrams for observed (OBS) and simulated soil water content (SWC) at 5 cm depth. Each marker shows one daily average. Top left diagram shows open loop (OL), top right shows assimilation of state variables (DA_s), bottom left shows data assimilation of state and parameters (DA_s+p), and bottom right shows data assimilation of state and parameters including organic matter (DA_s+p+o). Each diagram includes the root mean square error (RMSE), unbiased root mean square error (ubRMSE), mean bias error (MBE), and squared correlation coefficient (R^2).

https://doi.org/10.5194/gmd-2021-38
Preprint. Discussion started: 8 September 2021



535 **Figure 5: Correlation diagrams for observed (OBS) and simulated soil water content (SWC) at 20 cm depth. Each marker shows one daily average. Top left diagram shows open loop (OL), top right shows data assimilation of state variable (DA_s), bottom left shows data assimilation of state and parameters (DA_s+p), and bottom right shows data assimilation of state and parameters including organic matter (DA_s+p+o). Each diagram shows root mean square error (RMSE), unbiased root mean square error (ubRMSE), mean bias error (MBE), and squared correlation coefficient (R^2).**

https://doi.org/10.5194/gmd-2021-38
Preprint. Discussion started: 8 September 2021

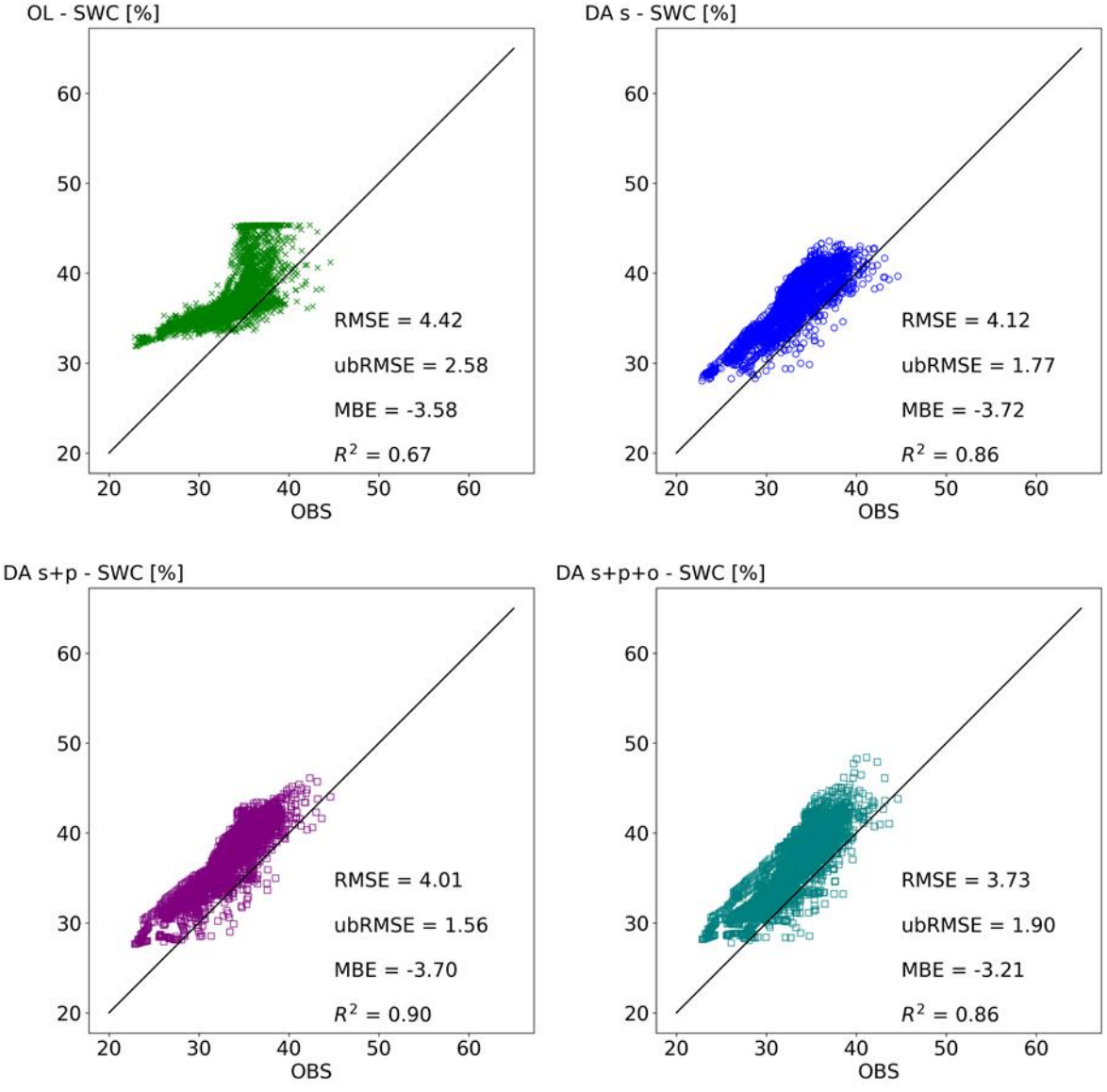


Figure 6: Correlation diagrams for observed (OBS) and simulated soil water content (SWC) at 50 cm depth. Each marker shows one daily average. Top left diagram shows open loop (OL), top right shows data assimilation of state variable (DA_s), bottom left shows data assimilation of state and parameters (DA_s+p), and bottom right shows data assimilation of state and parameters including organic matter (DA_s+p+o). Each diagram shows root mean square error (RMSE), unbiased root mean square error (ubRMSE), mean bias error (MBE), and squared correlation coefficient (R^2).

Table 1: Statistical properties and cross-correlation coefficients (CC) used to perturb the atmospheric forcing data.

https://doi.org/10.5194/gmd-2021-38
Preprint. Discussion started: 8 September 2021


Geoscientific Model Development Discussions

	Perturbation	Mean	Standard deviation	CC **PR**	CC **SW**	CC **LW**	CC **TP**
Precipitation (**PR**)	Multiplicative log-normal distribution	1.0	0.5	1.0	-0.8	0.5	0.0
Shortwave radiation (**SW**)	Multiplicative log-normal distribution	1.0	0.3	-0.8	1.0	-0.5	0.4
Longwave radiation (**LW**)	Additive normal distribution	0.0	20.0	0.5	-0.5	1.0	0.4
2m Air temperature (**TP**)	Additive normal distribution	0.0	1.0	0.0	0.4	0.4	1.0

Table 2: Statistical evaluation measures for the four different simulation and assimilation scenarios, always compared to measurements.

	OL	DA_s	DA_s+p	DA_s+p+o
RMSE / 5cm	8.08	5.59	5.24	4.82
ubRMSE / 5cm	6.52	4.19	3.29	3.47
MBE / 5cm	-4.78	-3.7	-4.07	-3.35
R^2 / 5cm	0.63	0.91	0.93	0.92
RMSE / 20cm	4.03	2.84	2.27	1.67
ubRMSE / 20cm	3.56	1.62	1.86	1.64
MBE / 20cm	1.89	2.33	1.3	0.31
R^2 / 20cm	0.66	0.96	0.92	0.95
RMSE / 50cm	4.42	4.12	4.01	3.73
ubRMSE / 50cm	2.58	1.77	1.56	1.9
MBE / 50cm	-3.58	-3.72	-3.7	-3.21
R^2 / 50cm	0.67	0.86	0.9	0.86

Table 3: Initial soil texture data and soil texture data after updating by data assimilation.

Type / depth	Initial ensemble mean	Updated ensemble mean	Updated ensemble standard deviation

https://doi.org/10.5194/gmd-2021-38
Preprint. Discussion started: 8 September 2021


Geoscientific Model Development Discussions

Sand / 5 cm	19.3	45.7	13.0
Sand / 20 cm	23.3	49.1	12.3
Sand / 50 cm	27.3	52.6	11.3
Clay / 5 cm	38.9	35.0	12.2
Clay / 20 cm	38.9	34.9	10.9
Clay / 50cm	37.9	33.4	10.5
Organic matter / 5 cm	34.1	51.4	8.17
Organic matter / 20 cm	15.8	32.3	7.8
Organic matter / 50 cm	8.7	13.1	4.9