LACE
nwp central europe

Regional Cooperation for Limited Area Modelling in Central Europe

RC LACE Stay Report

Topic: Assimilation of Satellite Based SWI Observations using a Simplified Extended Kalman Filter Algorithm in SURFEX

OMSZ Budapest 30th May - 24th June 2022

Matjaž Ličar

Supervisor: Helga Tóth

CONTENTS

1 Introduction

Surface processes have an impact on the state of the lower atmosphere. It is thus desirable to provide accurate initial conditions to land surface models in a NWP system. Data assimilation methods provide us with means to achieve this. Different observation types can be used. For data assimilation in land surface models, the in-situ observations of surface quantities have the disadvantage that correlation drops quickly with distance from the measurement location. An advantage of remote sensing data is, that it typically covers a large portion of the domain of interest.

Recently, Helga Tóth and Balázs Szintai from the Hungarian meteorological service have performed data assimilation using a simplified extended Kalman filter algorithm of a satellite based LAI product into the SURFEX land surface model [1]. The objective of this stay was to assess the feasibility from a technical perspective to assimilate, in a similar fashion, a different soil moisture satellite product corresponding to the modeled superficial soil moisture.

2 Copernicus Soil Water Index Satellite Product

The Copernicus Global Land Service offers a wide variety of satellite observation based land surface products at different spatial and temporal resolutions. In our experiment we used their Soil Water Index product [2] as an observational measure of water content in the superficial soil layer, to be assimilated in the land surface model.

This product is a fusion of two existing surface soil moisture (SSM) products from different sensors. For Sentinel-1, this is produced via C-band SAR imaging at a high spatial resolution of 1km. The temporal availability spans a period between 1.5 to 4 days. SSM products are also produced with the Metop ASCAT sensor at a coarser spatial resolution of 0.1°, however it is available on a daily basis. By combining both, we obtain a product which achieves both high spatial and temporal resolution.

Equation (1) describes, how SSM moisture measurements from both sensors are transformed into SWI:

$$\text{SWI}_T^w(t_n) = \frac{(n-1+1)\sum_i^n w(t_i)\text{SSM}(t_i)e^{-\frac{t_n-t_i}{T}}}{\sum_i^n w(t_i)e^{-\frac{t_n-t_i}{T}}} \quad \text{for } t_i \leq t_n, \tag{1}$$

where t_n is the observation time of the current measurement, t_i are the observation times of previous measurements (both in units of days), factor T determines how strongly past SSM observations influence the current SWI and w are weights for each sensor. In essence, this formulation represents a simple two-layer water balance model, where the superficial soil layer can be observed by satellite sensors, and is connected to the second layer as the part of the profile which extends downwards from the bottom of the soil surface layer. Temporal variability decreases with increasing profile depth, which means a higher value of T corresponds to deeper soil layers. Since we are interested in the superficial layer in our experiment, we take T equal to 1, however data for other values are provided in the product as well. Weights w are determined on the basis of signal to noise ratio of both sensors.

3 Bias Correction Using a CDF Matching Method

Similarly to any observations, our satellite product data set also contains biases and various differences with data present in the land surface model. Satellite observation data thus needs to be calibrated before data assimilation can be performed. For this purpose, we use a cumulative distribution function (CDF) matching method [3]. This method is useful when we wish to remove systematic differences and to rescale data obtained from one source to another. This method requires long enough data sets from both sources. In our case, we used daily satellite SWI data for the period from May 2021 to May 2022, and a model data set from operational AROME analyses for the same time period. A CDF is then computed for each grid point and both data sets. The basis of this method is then the assumption, that for a given observation value, the values of the CDF functions of the model and observation data sets will be equal:

$$\text{CDF}(x_s) = \text{CDF}(x_m), \tag{2}$$

where x_s denotes the observation value and x_m the model equivalent. In principle, x_m and x_s do not need to be of the same units. Assuming both data sets are distributed with a Gaussian distribution with a certain mean μ and standard deviation σ, a simple solution can be obtained by solving equation (2) analytically. We use the known expression of a CDF of such a distribution in equation (2):

$$\frac{1}{2}\Big[1+\mathrm{erf}\Big(\frac{x_s-\mu_s}{\sqrt{2}\sigma_s}\Big)\Big]=\frac{1}{2}\Big[1+\mathrm{erf}\Big(\frac{x_m-\mu_m}{\sqrt{2}\sigma_m}\Big)\Big]. \tag{3}$$

For this expression to hold, the arguments of both Gaussian error function arguments need to be equal. We thus obtain

$$\begin{aligned} x_m &= p_2 x_s + p_1 \\ p_2 &= \frac{\sigma_m}{\sigma_s} \\ p_1 &= \mu_m - p_2\mu_s \end{aligned} \tag{4}$$

To perform the calibration using expression (4), we only need to compute the mean and standard deviation of both data sets for each grid point. This expression was also used to perform the calibration in our assignment.

Alternatively, we can find the solution to equation (2) numerically. Many high level programming languages such as **R** have a fast performing built in function to compute the CDF of a given data set. A simple polynomial regression analysis of such a CDF can then be performed with equal simplicity. The root of this polynomial can then be calculated with Newton iteration, which converges in only a couple of steps and works robustly for our case. We can then compute the roots spanning the entire range of SWI (0 to 100) and use polynomial regression once again to obtain a nonlinear relationship between satellite measurements and modeled soil moisture. Such coefficients can then be stored in look up tables for efficient operational use. This method is demonstrated on figure (1).

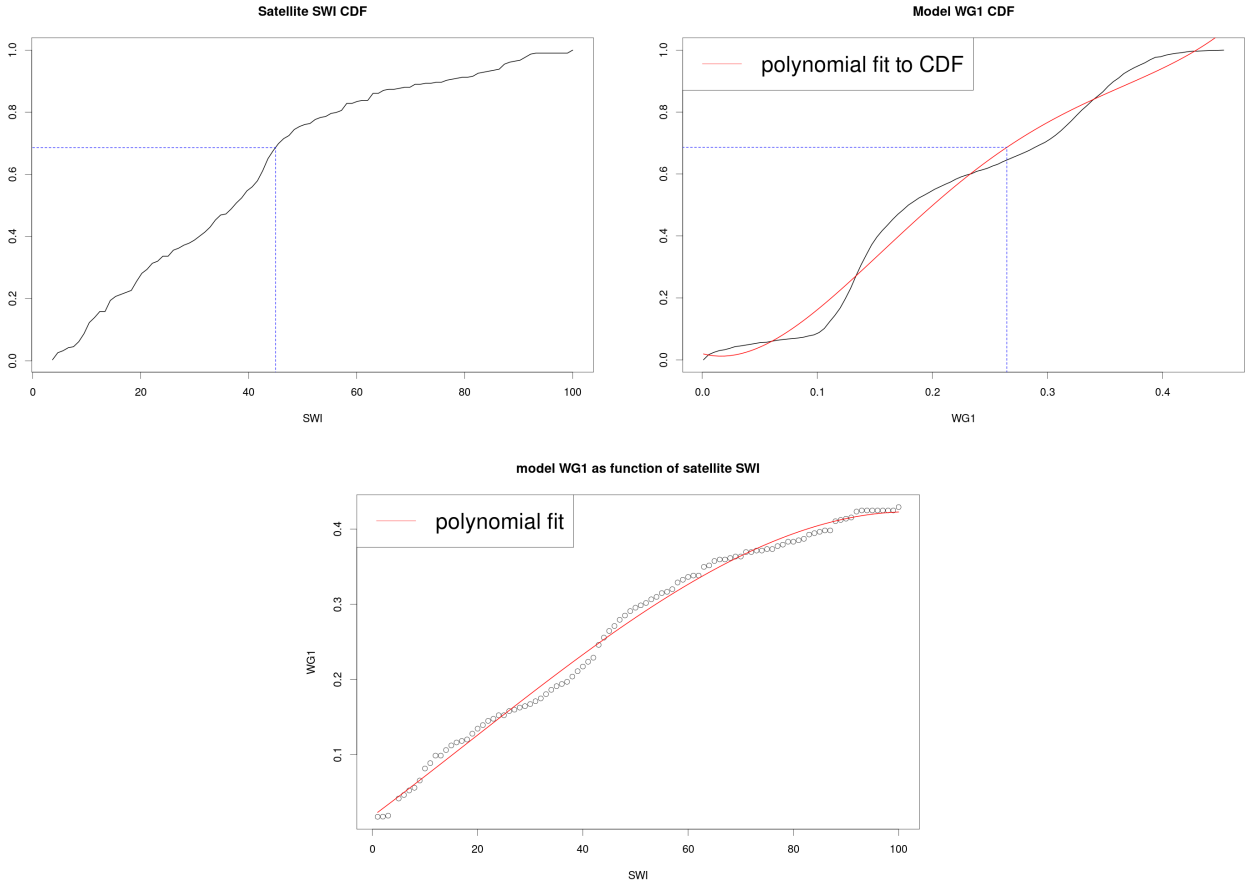

FIGURE 1. Calibration of satellite data using CDF matching. Top left: CDF of satellite SWI observations. Top right: CDF of the modeled superficial soil layer WG1 with a third order polynomial fit. Bottom: Modelled WG1 as a function of observed SWI and a third order polynomial fit.

Proper validation is needed to see if this method yields any benefits compared to the linear analytical expression (4). The determination of the most appropriate polynomial order or the functions used in regression analysis could also be investigated further. This is however beyond the scope of our topic.

4 Preprocessing of Satellite Data

Satellite data is provided on 1km latitude and longitude grid which spans the entire Europe and a small portion of surrounding regions. These data were obtained by the Hungarian NWP team beforehand and is stored in netcdf format. The archive spans the time period between May 2021 and May 2022, with each daily file containing data valid at 12UTC. These data need to be interpolated to the land surface model grid, rescaled and treated for biases with erroneous data removed and stored in an appropriate format, before it can be used in our data assimilation experiment.

Bad data is removed in the first step. Here we use the gridded surface state flag and the quality flag values also available with the satellite product. The land flag indicates frozen/unfrozen/melting state of the surface at the grid points, while the quality flag indicates grid points with a very poor signal to noise ratio of the data. These data are then discarded. Climate Data Operators (CDO)[4] is a command line utility which is able to perform manipulation of data defined on various geographic grids and supports various file formats. In the second step we use this utility to interpolate the satellite data to the domain of the operational AROME NWP model, which is a 2.5km Lambert grid centered on Hungary, as it will also be the domain used in our experiment. Finally, the data are calibrated according to the process described in section (3) and stored in ASCII format (one value per line, one file per date/time). These steps are shown on figure (2).

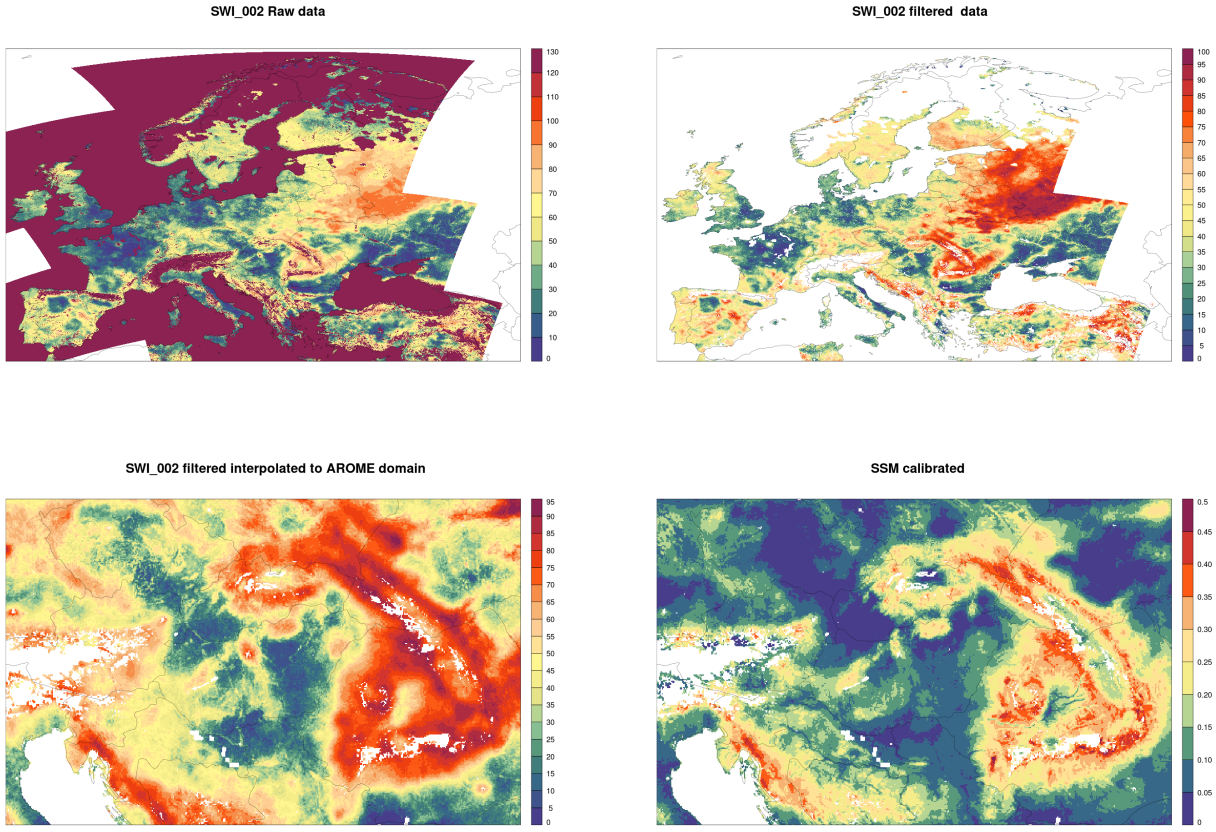

FIGURE 2. Preprocessing of satellite data. Top left: Raw SWI data from the satellite product. Top right: Raw SWI data with erroneous data discarded. Bottom left: SWI data interpolated to the AROME domain. Bottom right: Bias corrected SWI data using the CDF matching method.

5 SURFEX PLATFORM AND THE ISBA LAND SURFACE MODEL

SURFEX is a land surface modelling platform [5]. Essentially, it is a collection of different models of the earths surface, with a common interface for initialization, definition of physiographic data and treatment of boundary conditions. We can choose different schemes depending on the surface type. These schemes are one dimensional column models, so the model for each grid point runs independently of the others. SURFEX can be set up to run on

a variety of regular grids with a broad range of spatial resolutions and has a wide range of applications, including climatology, data assimilation and operational NWP. A time series of atmospheric boundary conditions can be prescribed for a run in offline mode. Direct coupling with an atmospheric model is also possible and is implemented for several models.

The ISBA (Interaction Sol-Biosphère-Atmosphère) scheme is used to model the vegetation and bare soil land surface in SURFEX. It accounts for heat and mass exchanges between the atmosphere and the soil up to a certain depth. Several options for the treatment of soil and vegetation physics, as well as snow pack sub models can be selected. In our experiment we use the 3 layer force restore option.

The 3 layer force restore option for ISBA [6] uses two prognostic variables for the temperature. The prognostic equation for the surface temperature $TG1$ accounts for the thermodynamic fluxes at the boundary between the soil and the atmosphere and a relaxation term to the bulk soil temperature $TG2$, which is governed by a second prognostic equation. Soil moisture is treated in a similar fashion with prognostic equations for the superficial soil layer $WG1$, the root zone layer $WG2$, and additionally for the 3 layer option the base layer $WG3$. The exchange of moisture for $WG3$ occurs only with $WG2$ and the atmosphere through evapotranspiration. Phase change of soil moisture is also possible, which is accounted for by additional soil ice reservoirs in the ISBA model.

The version of SURFEX used for our experiment is compiled from the operational Hungarian AROME cy43 code, which uses SURFEX v8.0 and has some minor modifications implemented by the Hungarian NWP team.

6 Simplified Extended Kalman Filter in SODA

Data assimilation when running SURFEX in offline mode is performed by running the SODA (SURFEX offline data assimilation) binary. Several data assimilation algorithms are implemented, including optimal interpolation (OI), extended Kalman filter (EKF) and ensemble Kalman filter (EnKF). As is the case with the model itself, the data assimilation schemes

in SODA also operate point wise. As a consequence, the model state vector has a substantially lower dimension compared to upper air data analysis and the assimilation equation can be solved explicitly.

The analysis update equation is given by (5) [5, 7]:

$$x_a(t_i) = x_f(t_i) + K_i(y_o(t_i) - h_i[x_f]), \tag{5}$$

where x denotes the model state vector with the subscripts a denoting the analysis, f the existing forecast, y_o the vector of observations, and h_i the observation operator. The Kalman gain matrix K_i is given by (6):

$$K_i = BH^T(HBH^T + R)^{-1}, \tag{6}$$

where B is the background x_f error covariance matrix, H is the linearized observation operator h_i in matrix form and R is the observation error covariance matrix. The linearization of the observation operator is generally referred to as the extended Kalman filter. When running EKF in SODA, the linearized observation operator is estimated by applying small perturbations of the components of the control vector x_j:

$$H_{mj} = \frac{\partial y_m}{\partial x_j} \approx \frac{y_m(x + \delta x_j) - y_m}{\delta x_j}, \tag{7}$$

where δx_j is the user prescribed perturbation magnitude.

In principle, the background covariance matrix B evolves in time. Such computations can be very time consuming, so a common simplification is to keep B as constant. This assumption is generally referred to as the simplified extended Kalman filter.

7 Experiment Setup

SURFEX runs are performed in three steps, each by running its corresponding binary. First, the PGD binary is run to produce a physiographic PGD file containing the information

about the grid, orography, the various soil and vegetation parameters used by ISBA, the surface schemes and their configuration used by SURFEX. The domain for our experiment is identical to the one used for the operational Hungarian AROME model so we do not need to produce the PGD files ourselves and can use the operational one instead.

The initial conditions are defined by running the PREP binary. This initializes the prognostic variables in all the surface schemes to prescribed values. Initial conditions can be read from files of various formats, or they can be set to uniform values for the entire grid. The most convenient for us is to use the existing AROME analyses mentioned earlier, as SURFEX is able to produce PREP files, which store initial conditions, from such a file type directly.

Model integration is performed with the OFFLINE binary. To perform this, we need the PGD and PREP files mentioned above, and a FORCING file, which contains the atmospheric boundary conditions for our experiment. This file is typically in netCDF format, although the use of ASCII files is also possible. The format along with correct field names for all the required atmospheric boundary conditions of such a netCDF file is described in the documentation [8]. Such forcing files for the 6 month period between May 2021 and November 2021 are available in the archive of the Hungarian NWP service, so we do not need to produce our own. Each of these files contains the atmospheric boundary conditions from AROME analyses for 24 hours at an hourly time step. Running the integration produces selected output at a selected time steps, as well as an output file containing the entire model state, which is identical in format to the PREP file. This file can then be used as the initial conditions for a subsequent integration.

These steps are required to perform an experiment with no data assimilation. To asses the effects of data assimilation, we perform a control run in this manner.

We use a 24 hour assimilation window for our assimilation experiment, since we are only interested in one observation time which is available at this frequency. To perform data assimilation, several different files need to be provided. First, we compute the model background. This is performed as a 24 hour run in the same fashion as described above, where we take the initial state from the previous analysis. This file needs to be named as `PREP_INIT.fa`.

Perturbed runs to estimate the Jacobian are also required when using SEKF. A seperate perturbed run for each of the control variables is needed. We perform the perturbed runs by setting the appropriate namelists in `OPTIONS.nam` file. We can also specify which model variables we wish to treat as control variables. We can then set up different experiments differing in the number control variables. The nature of our observations presents us with several different cases to try, for example using only the superficial soil moisture layers $WG1$ and $WG2$, or treating the soil temperatures $TG1$ and $TG2$ in addition to the soil moisture as the control variables, to better learn about the effects of data assimilation. The names of the output of the perturbed runs are hardcoded as well and need to follow the naming convention `PREP_${YY}${MM}${DD}H${hh}_EKF_PERT${i}.fa`, where i refers to the control variable.

Observations need to be provided in an ASCII file following the naming convention `OBSERVATIONS_${YY}${MM}${DD}${hh}.DAT`, where we put one observation value on separate lines for each grid point. The input of $WG1$ observation type was not yet implemented in the version of SURFEX used in our experiment, so a trivial modification of the `soda.F90` routine, highlighted on figure (3), was necessary. If we were to setup a more complicated experiment involving the assimilation of multiple observation types using SEKF, much additional work in terms of modifying the SURFEX code would be needed.

FIGURE 3. Modification in the `soda.F90` routine: left new version, right old version.

Once the reference and the perturbed runs are completed and the observations are available we perform the data assimilation by running the SODA binary. This produces a file

```
&NAM_ASSIM                    !General assimilation namelist
  LASSIM = T,                 !To perform DA
  CASSIM_ISBA = "EKF "        !Choose DA algorithm (also OI,ENKF)
/
&NAM_OBS                      !Observations namelist
  NOBSTYPE = 1,               !Number of different observations
  CFILE_FORMAT_OBS = "ASCII",!File format for observations
  XERROBS_M = 0, 0, 0.1,      !Observational standard deviation
  NNCO = 0, 0, 1,0,0,         !Vector of observation types to be assimilated
/                             !order: "T2M","HU2M","WG1","LAI","SWE"
&NAM_VAR                      !Namelist for control variables for EKF
  NVAR = 2,                   !number of control variables
  NIVAR = 1,                  !perturbation of i-th variable for perturbed run
  NNCV = 1,1,0,0,0,           !vector of control variable types (in correct order)
  XSIGMA_M = 0.1, 0.15,       !control variable standard deviations
  XTPRT_M = 0.0001,0.0001,    !perturbation amplitudes
/
&NAM_IO_VARASSIM              !General SODA EKF options
  LBEV = F,                   !Compute time evolution of B
  LBFIXED = T,                !Constant value of B
/
```

FIGURE 4. Assimilation related namelists in `OPTIONS.nam`, which need to be defined when running `OFFLINE` for perturbed runs and running `SODA` using EKF.

named `SURFOUT.${YYYY}${MM}${DD}_00h00.fa` containing the analysis. The assimilation related namelist settings in `OPTIONS.nam` are shown on figure (4) for reference.

8 RESULTS

In this section we show and comment some of the results from satellite data preprocessing and the data assimilation experiments. Three locations, referred to as P1, P2 and P3 were chosen to present the evolution of quantities in question with time at a specific location. Some basic consideration was made in the choice of these locations to have some differences in the physiographic parameters. The locations of these points are shown on figure (5).

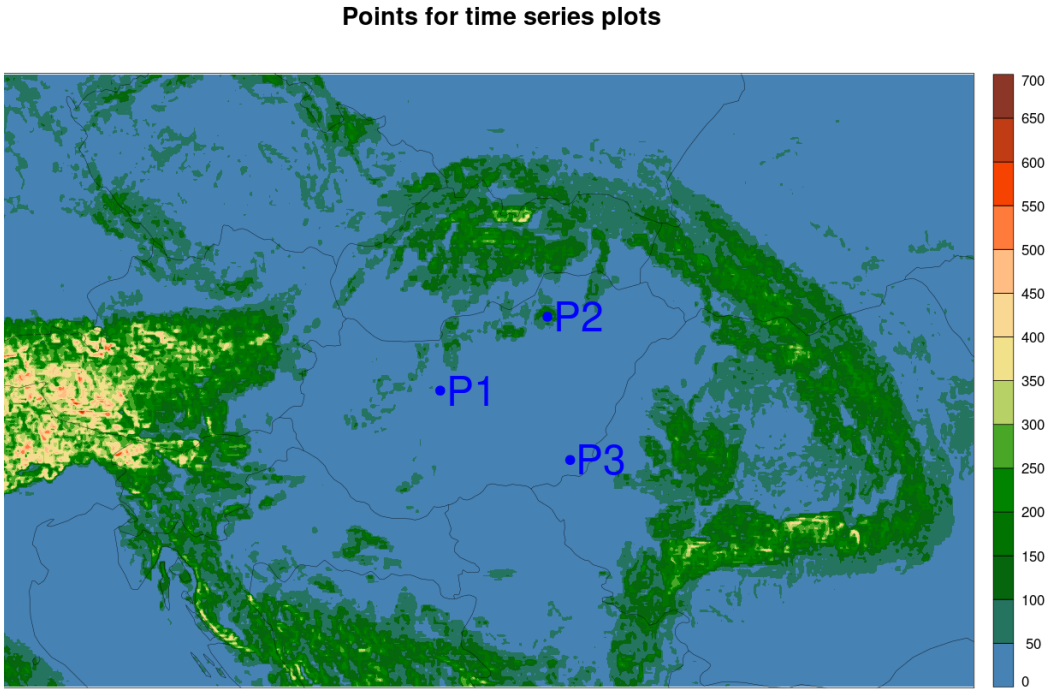

FIGURE 5. The location of points chosen to present data for this report.

8.1 Preprocessing of Satellite Observations. We assess the performance of the calibration process by comparing the raw SWI data from the satellite measured as a percentage, the rescaling of this quantity into the model units (specific water content in units of $[m^3/m^3]$),

both the analytic and numeric calibration method described earlier and the AROME analyses. The rescalation from SWI to model domain SSM is a simple linear transformation based on the minimum and maximum value in the AROME analysis dataset. Results are shown on figure (6).

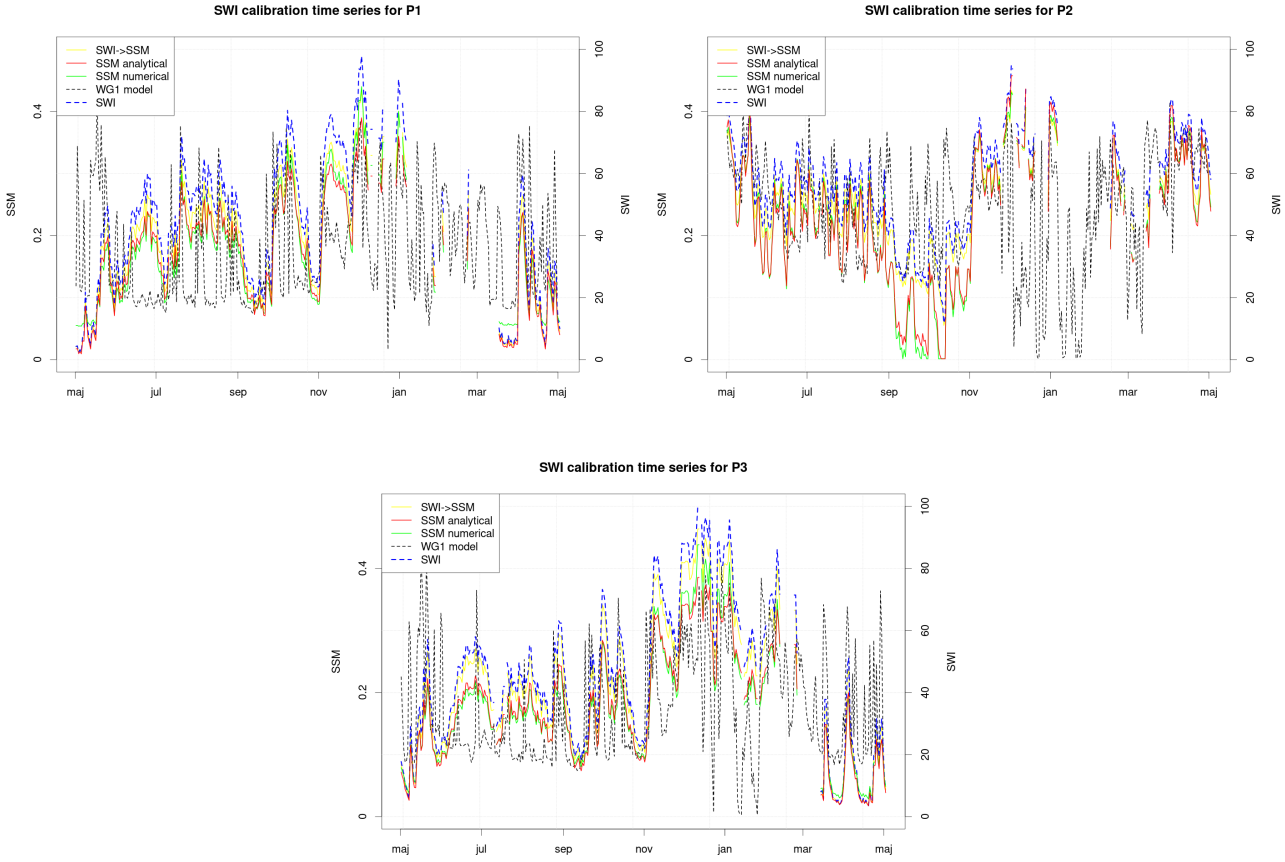

FIGURE 6. Soil moisture observation processing for points P1 (top left), P2 (top right) and P3 (bottom). Yellow: Linear rescaling from SWI (in percentage) to SSM (in $[m^3/m^3]$) Red: SSM calibration with the analytical method. Green: SSM calibration with the numerical method. Black: $WG1$ from AROME analyses. Blue: Raw satellite SWI data.

We see, that the calibration process slightly shifts the raw satellite data towards the AROME analyses. The difference between the analytic and numerical method is apparent, however small. It is thus questionable, if the little extra effort for the numerical method is justified. We should also note the large difference for P2 between September and November. Most likely, the difference arises due to the missing observation data between January and March. In this period, the AROME analysis shows very low values, which are implicitly included in the calibration process. This suggests, that longer data sets should be used and the model analysis data, where observation data is unavailable, should be discarded.

8.2 Data Assimilation Experiments. There are many possibilities to set up our assimilation system in terms of the control variables and user prescribed parameter values. We limit ourselves to 3 experiments, which differ in the number of control variables and the magnitude of the observation error σ_{obs}:

- Experiment 1: $\sigma_{obs} = 0.1$, control variables $WG1$, $WG2$
- Experiment 2: $\sigma_{obs} = 0.03$, control variables $WG1$, $WG2$
- Experiment 3: $\sigma_{obs} = 0.1$, control variables $WG1$, $WG2$, $TG1$, $TG2$

Initial plan was to run each of these experiments for the period of six months between May 2021 and November 2021, where the forcing files are available. Unfortunately, the initial output was clearly erroneous, the cause for which was not identified before the stay came to its conclusion. Afterwards, Helga Tóth managed to find an oversight in the experiment scripts and reran the experiments for a shorter period of 1 month. It is thus more appropriate to present her results as opposed to what was originally intended.

Figure (7) shows the satellite data along with the $WG1$ fields for each experiment valid for in the middle of the 1 month run, i.e. 16 May 2021. Experiments 1 and 3 are identical and show a certain degree of similarity with the observation, however the results of experiment 2 with a much lower σ_{obs} do not appear to be very convincing. Note also, that very little consideration has been given to the magnitudes of σ_{obs} for the experiments, as the goal was simply to observe the effects of choosing different values on the behaviour of the system. One

would intuitively expect that a lower σ_{obs} would put more emphasis on the observations and the analysis would thus be closer to the observations than for higher values.

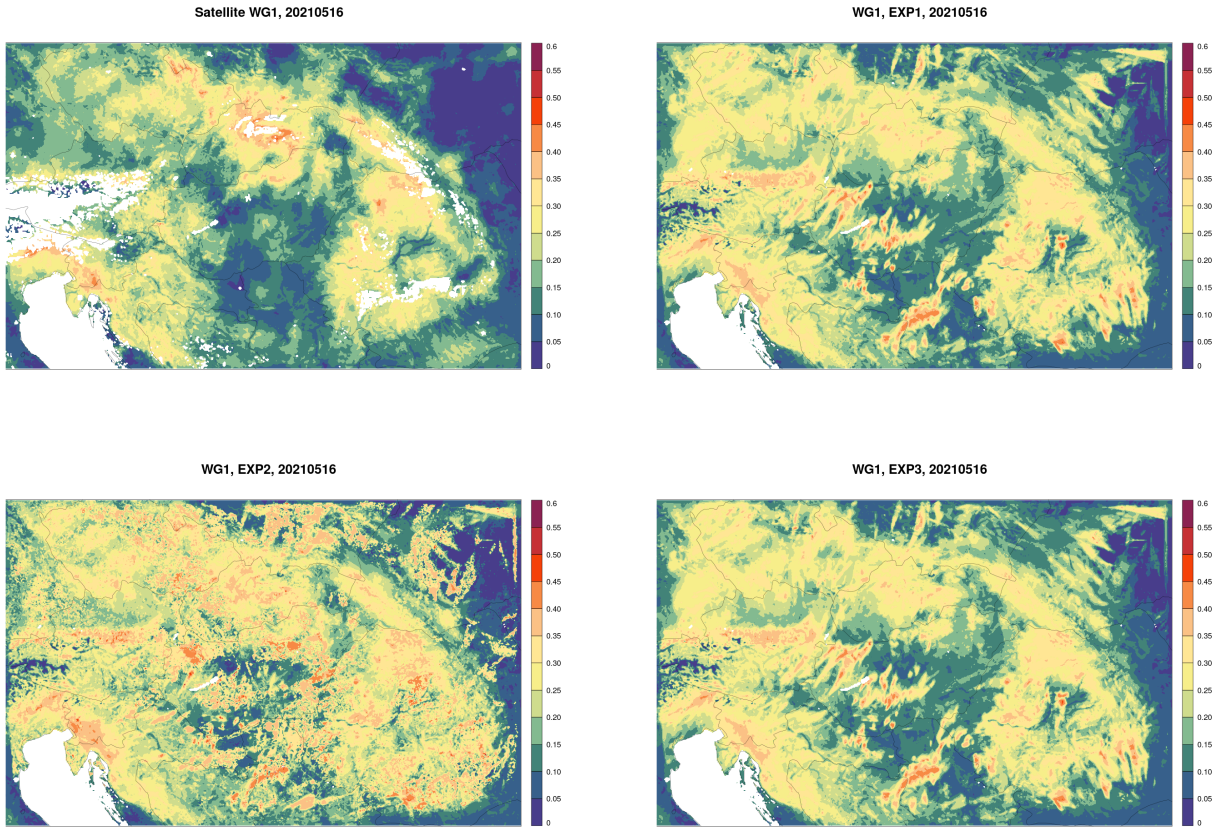

FIGURE 7. Comparison between the satellite data (top left) and the modeled superficial soil moisture $WG1$ for experiments 1 (top right), 2 (bottom left) and 3 (bottom right). Figures were produced by Helga Tóth.

Figure (8) shows the time series of observations and modeled quantities for P1. Compared to the reference open loop run without data assimilation, the identical results for $WG1$ in experiments 1 and 3 exhibit a slight decrease towards observations. The amplitude of the variability in experiment 2 is however much larger. We see a similar behaviour for $WG2$.

Interesting to note here is, that the open loop run remains nearly constant and that nearing the end of the 1 month period the results of experiment 1 and 3 start to differ.

Assimilation of observations of soil moisture have very little effect on superficial soil temperature $TG1$ compared to the control run. There is however a notable increase in $TG2$ for experiment 2.

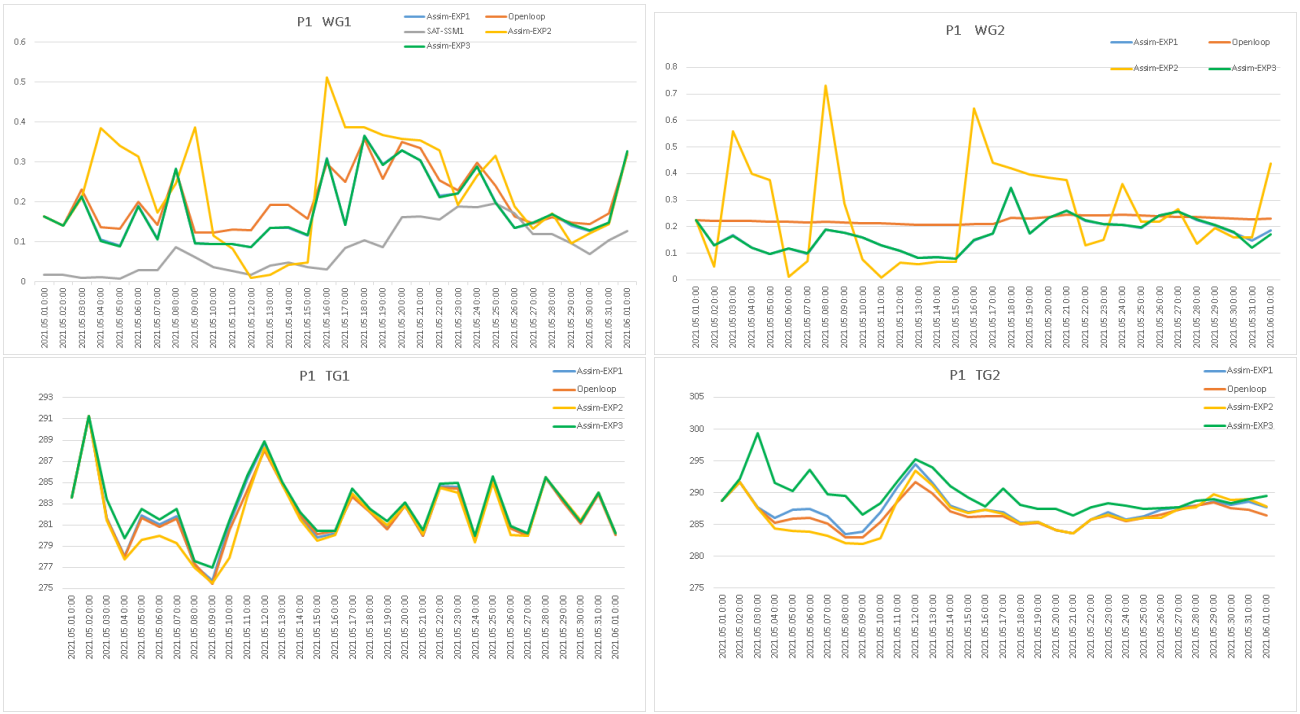

FIGURE 8. Time series of modeled quantities for the different experiments. Top left: $WG1$ with observations. Top right: $WG2$. Bottom left: $TG1$. Bottom right: $TG2$. Figures produced by Helga Tóth.

9 Conclusion

We have demonstrated the procedure how satellite based soil moisture can be assimilated using SEKF in SODA. The time available for the stay unfortunately did not allow for anything but the most basic results to be examined. It would certainly be beneficial to rerun the experiments for a longer time period and more cases with different values for observation and background errors. It would also be interesting to examine the effects of such an assimilation system on screen level humidity and temperature. Performing this would not yield much additional work and would hopefully give us some insight on what benefits to operational NWP the assimilation of satellite observations of soil moisture might be.

Ultimately, the goal would be to run such a system operationally. This presents several additional challenges. An operational system should account for many observation types, valid and available at different time scales. In principle, there are a number of possibilities on how to design such a system, so careful consideration is needed to choose the most appropriate option. This would potentially also require additional work on the SODA/EKF code to be more flexible in terms of supporting more simultaneous observation types, which the user might provide in separate files of different types.

We have treated the observation error as a user provided parameter in our experiments. A more appropriate estimation of these parameters is clearly needed.

We have used the 3 layer ISBA soil model in our experiments. In SURFEX we also have the option to use the multiple layer ISBA diffusion scheme. Our satellite product also contains data corresponding to deeper soil layers. This could also be a basis for future work. A similar exercise was recently performed by the Austrian NWP [9].

References

[1] Tóth, H., & Szintai, B. (2021). Assimilation of Leaf Area Index and Soil Water Index from Satellite Observations in a Land Surface Model in Hungary. Atmosphere, 12(8).

[2] Copernicus Global Land Operations. (2022). Vegetation and Energy CGLOPS-1 PRODUCT USER MANUAL, SOIL WATER INDEX, COLLECTION 1KM, VERSION 1.

[3] Reichle, R., & Koster R. (2004). Bias reduction in short records of satellite soil moisture. Geophysical Research Letters, 31(19).

[4] Uwe Schulzweida. (2021). CDO User Guide Version 2.0.5

[5] Le Moigne, P. (2018). SURFEX scientific documentation, 87, 211.

[6] Boone, J. C. Calvet, and J. Noilhan. (1999) Inclusion of a third soil layer in a land surface scheme using the force-restore method. J. Appl. Meteorol., 38(11):1611–1630

[7] De Rosnay, P., Drusch, M., Vasiljevic, D., Balsamo, G., Albergel, C., & Isaksen, L. (2013). A simplified Extended Kalman Filter for the global operational soil moisture analysis at ECMWF. Quarterly Journal of the Royal Meteorological Society, 139(674), 1199-1213.

[8] SURFEX. n.d. SURFEX homepage. (2022). www.umr-cnrm.fr. https://www.umr-cnrm.fr/surfex/ .

[9] Vural, J., Schneider, S., Bauer-Marschallinger, B., & Haslinger, K. (2020). Assimilation of the SCATSAR-SWI with SURFEX: Impact of local observation errors in Austria. Monthly Weather Review, 149.