CopERnIcus climate change Service Evolution

D1.3 Unified, ensemble-based regional land data assimilation system

Due date of deliverable	31st December 2024
Submission date	8 January 2025
File Name	CERISE-D1-3-V1.1
Work Package /Task	WP1 Tasks 1.2 & 1.3
Organisation Responsible of Deliverable	SMHI
Author name(s)	Jelena Bojarova, Abhishek Lodh, Jostein Blyverket, Åsmund Bakketun, Jude Musuuza, Ilaria Clemenzi, David Gustafsson, Tomas Landelius
Revision number	V1.1
Status	Issued
Dissemination Level	PU

Funded by the European Union

The CERISE project (grant agreement No 101082139) is funded by the European Union.
Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the Commission. Neither the European Union nor the granting authority can be held responsible for them.

1 Executive Summary

This report summarizes the development done in the CERISE project towards unified ensemble-based data assimilation framework for regional reanalysis applications. Regional reanalysis applications require an adequate treatment of soil and surface variables on a wide range temporal and spatial scales. The CERISE project aims to progress on three following topics:

- A homogenization of the analysis of snow and soil variables is one the aims of the CERISE project. At present different model variables are analysed using different methods, which is leading to the inconsistencies in the analysis and to heavy maintenance burden.
- A development of flexible ensemble-based data assimilation into the ISBA-Diff soil model and the multi-layer snow scheme is the second task of CERISE. Advanced physical models together with flexible data assimilation schemes able to handle a variety of observations from different platforms will improve quality reanalysis products of near surface variables.
- First steps towards development of a consistent hydrological - meteorological forecasting system that is necessary to properly address evolution of water cycle including snow are taken in CERISE.

The work has been carried out in three different teams with the aim of integrating the successful outcomes of the development into the common CERISE HARMONIE-AROME code release. Ensemble Kalman Filter was chosen as a unified ensemble-based land data assimilation framework. The Meteorology Research team at SMHI has been focusing on the development of the Ensemble Square-Root Kalman Filter (EnSRKF) data assimilation scheme in the inline HARMONIE-AROME environment. An extensive comparison to the Simplified Extended Kalman Filter (sEKF) scheme, used as a reference scheme, in the HARMONIE-AROME CY46 has been performed. The EnSRKF filter propagates the information from the screen level variables deeper into the soil and provides the analysis that better agrees with observations. At the same time the EnSRKF provides a somewhat too strong response to the daily variations caused by the daily cycle. The reason for this behavior and the possible remedies are under investigation. The Met Norway team has been focusing on the development of the Local Ensemble Transform Kalman Filter (LETKF) scheme, which is a variant of EnSRKF. The scheme has initially been developed for the snow data assimilation in the off-line environment and has been extended to the analysis of soil variables. Both EnSRKF and LETKF schemes have a similar rescaling engine but differ by way observations are handled. The LETKF scheme has an attractive possibility to treat a footprint of the satellite instrument in an explicit way. The LETKF scheme is able to significantly improve the estimate of snowpack in the areas of flat orography. The LETKF scheme has been evaluated as well in the inline environment for analysis of soil variables using screen level observations. The preliminary evaluation of the Kalman Gain shows reasonable and intuitively sensible results. The Hydrology Research Team at SMHI has been involved in the development of the Ensemble Kalman Filter (EnKF) for snow data assimilation. The focus was here on adequate treatment of the snow memory that can be difficult to capture for the timescales of the NWP model. The sensitivity of the EnKF to tunable parameters, such as observation error variance and horizontal and vertical localisation length scales, have been investigated. Further development of the hydrologically constrained localisation for assimilation of upstream observations is under implementation.

The ultimate goal of the investigations carried out within this deliverable “unified ensemble-based regional land data assimilation system” is to provide a solid foundation for the next generation of Copernicus reanalysis in a form of a robust land surface data assimilation. Three teams are involved in the developments approach this goal from different perspectives. The EnSRKF development is carried out from the starting point in the efficient functionality of the scheme within the in-line Harmonie-AROME environment and investigation of impact of the ISBA-diff modelling on fluxes. The LETKF scheme focuses on the stand-alone environment,

such as required for CARRA-Land prototypes, and enabling of use of remote sensing observations, including modeling of the footprint-operator. Hydrological modelling will provide a framework to validate CERISE snow reanalysis products provided by the EnSRKF or LETKF against such accumulated quantities as river discharge observations, and in this way evaluate ability of the reanalysis products to capture spatial and temporal scales in an adequate way.

Table of Contents

Contents

2 Introduction

The aim of this deliverable is homogenisation of the land surface data assimilation methodologies for reanalysis. Due to historical reasons different land modelling components are being analysed separately that often leads to inconsistent initialization of different land variables. Homogenisation of land data assimilation schemes is a necessary step towards consistent initialisation of land variables from a variety of available sensors that observe the Earth system. An adequate description of the water cycle on a range of temporal and spatial scales requires a consistent hydrological meteorological modelling. Although this goal is outside the scope of CERISE project, in this deliverable we take the first step in that direction and implement data assimilation for meteorological model and for hydrological model in the same unified ensemble-based framework.

2.1 Background

The scope of CERISE is to enhance the quality of the C3S reanalysis and seasonal forecast portfolio, with a focus on land-atmosphere coupling.

It will support the evolution of C3S, over the project's 4 year timescale and beyond, by improving the C3S climate reanalysis and the seasonal prediction systems and products towards enhanced integrity and coherence of the C3S Earth system Essential Climate Variables.

CERISE will develop new and innovative ensemble-based coupled land-atmosphere data assimilation approaches and land surface initialisation techniques to pave the way for the next generations of the C3S reanalysis and seasonal prediction systems.

These developments will be combined with innovative work on observation operator developments integrating Artificial Intelligence (AI) to ensure optimal data fusion fully integrated in coupled assimilation systems. They will drastically enhance the exploitation of past, current, and future Earth system observations over land surfaces, including from the Copernicus Sentinels and from the European Space Agency (ESA) Earth Explorer missions, moving towards an all-sky and all-surface approach. For example, land observations can simultaneously improve the representation and prediction of land and atmosphere and provide additional benefits through the coupling feedback mechanisms. Using an ensemble-based approach will improve uncertainty estimates over land and lowest atmospheric levels.

By improving coupled land-atmosphere assimilation methods, land surface evolution, and satellite data exploitation, R&I inputs from CERISE will improve the representation of long-term trends and regional extremes in the C3S reanalysis and seasonal prediction systems.

In addition, CERISE will provide the proof of concept to demonstrate the feasibility of the integration of the developed approaches in the core C3S (operational Service), with the delivery of reanalysis prototype datasets (demonstrated in pre-operational environment), and seasonal prediction demonstrator datasets (demonstrated in relevant environment).

CERISE will improve the quality and consistency of the C3S reanalysis systems and of the components of the seasonal prediction multi-system, directly addressing the evolving user needs for improved and more consistent C3S Earth system products.

2.2 Scope of this deliverable

2.2.1 Objectives of this deliverables

This deliverable is a description of the components of the regional data assimilation system for the ISBA-diffusion soil model and screen level variables including multi-layer snow scheme implemented in the unified data assimilation framework.

2.2.2 Work performed in this deliverable

In this deliverable the work outlined in task 1.3 (*Unified Land Data Assimilation development)* and task 1.2 (*Development of ensemble-based land data assimilation approach for soil moisture*) is summarised. We describe here the methodology for the sequential initialisation of snow and soil variables from observations in three different frameworks: land data assimilation performed in the HARMONIE-AROME environment, land data assimilation in the stand-alone offline environment and the land data assimilation in the hydrological HYPE model. Even if the ultimate goal would be a development of the consistent meteorological-hydrological data assimilation that is necessary to obtain an adequate description of the water cycle, this goal is outside the scope of the CERISE project. Our aim here is to create a stable prerequisite for integration of the hydrological and meteorological data assimilation together at a later stage by implementing the development in a unified land data assimilation framework. Different flavors of the Ensemble Kalman Filter are applied for different applications.

2.2.3 Deviations and counter measures

No deviations have been encountered.

2.2.4 Reference Documents

[1] Project 101082139- CERISE-HORIZON-CL4-2021-SPACE-01 Grant Agreement

2.2.5 CERISE Project Partners

ECMWF	European Centre for Medium-Range Weather Forecasts
Met Norway	Norwegian Meteorological Institute
SMHI	Swedish Meteorological and Hydrological Institute
MF	Météo-France
DWD	Deutscher Wetterdienst
CMCC	Euro-Mediterranean Center on Climate Change
BSC	Barcelona Supercomputing Centre
DMI	Danish Meteorological Institute
Estellus	Estellus
IPMA	Portuguese Institute for Sea and Atmosphere
NILU	Norwegian Institute for Air Research
MetO	Met Office

3 Unified Ensemble-based Regional Land Data Assimilation system

In ensemble-based land data assimilation (LDA) the usage of the model is very important for at least the fact that the forecast model, in addition to propagating information forward in time, it also amplifies the instabilities and perturbations along some directions and damping in other directions. The sEKF (simplified Extended Kalman Filter) and EnKF (Ensemble Kalman Filter) are two ensemble data assimilation methods investigated in this deliverable. The major drawback of the simplified Extended Kalman filter (sEKF) based LDA is the non-linearity problem, which makes updating of error covariance matrix, by integrating the linearized forward model suboptimal. This difficulty is overcome by using the ensemble Kalman filter (EnKF). The EnKF is essentially a Monte-Carlo based approach that relies on generating an ensemble of N ensemble members of background (or forecast) states (Evensen 1994). In traditional EnKF when random perturbations are added to observations to compute the ensemble of background fields spurious correlations arise between the background and observation errors. This drawback has initiated the development of different flavors of EnKF data assimilation schemes, such as Ensemble square root Kalman filter (EnSRKF). EnSRKF does not perturb the observations leading to reduced sampling error in the analysis step; it updates the ensemble mean and deviations deterministically. EnSRKF lowers sampling noise in the analysis by avoiding the addition of noise to the observations during the update stage. The EnSRKF often performs better with smaller ensembles because it lowers random errors caused by observation disturbance, improving analysis accuracy. The ground surface contains highly nonlinear processes, such as soil moisture and vegetation dynamics, which benefit from more precise state estimation. EnSRKF, by lowering noise and increasing ensemble representation, frequently delivers more accurate estimates of land surface conditions such as soil moisture, temperature, and fluxes.

3.1 Ensemble-based filter for meteorological application

The advantage of using Ensemble Kalman Filter (EnKF) based algorithm for land data assimilation lies in its ability to effectively handle nonlinear land surface models and complex surface dynamics with simpler implementation. Theoretically the sEKF technique struggles with the evolution of the forecast error covariance matrix, leading to inaccuracies, whereas the EnKF method circumvents these issues by using an ensemble of model states, which avoids the direct computation of the error covariance matrix. In EnKF, the forecast error covariance is implicitly represented by the spread of the ensemble members, while sEKF method requires explicit computation of model Jacobians and update of the B-matrix. The EnKF uses the ensemble mean, which better represents the expected state, especially for both Gaussian and non-Gaussian distributions. The EnKF introduces uncertainty through stochastic model dynamics and physics when integrating each ensemble member, providing a more natural representation of uncertainty compared to sEKF's reliance on fixed error matrices. Thus, EnKF typically performs better in practical scenarios due to its flexibility in handling model nonlinearities, making it more robust than the sEKF in many real-world applications.

In numerical weather prediction and data assimilation, using the HARMONIE-AROME model we are in the process of developing the land data assimilation system (LDAS) using 'inline' and 'offline' systems to address the different requirements for different applications. The online LDAS systems are integrated and run concurrently with full scale atmospheric HARMONIE-AROME forecast or NWP model, directly assimilating observations to update the model state in real-time, constituting the systems coupled through the forecast model propagation. Such a system assures continuous exchange of information between atmospheric and land surface fluxes. The implementation of such systems requires technical solutions for parallelization that are crucial for tasks which are time-sensitive such as operational forecasting. On the other hand, offline LDAS systems have the flexibility to operate independently of the real-time heavy

atmospheric model forecasts. They can be typically used for retrospective analyses, research, and climatological studies, where the focus is on the evaluation of long-term trends, case studies, and importantly for the model improvements. For instance, offline assimilation can involve extensive ensemble simulations that have the scope for the development of the EnKF or advanced machine learning models in a cost-efficient way. By developing and researching on separate LDAS systems, we can optimize our resources, balancing the needs of operational NWP forecasting with the longer-term goals of model development.

3.1.1 Ensemble Kalman Filter in the HARMONIE-AROME environment

The initialization of atmospheric NWP models with accurate land surface state coupled with land-surface/vegetation models is a key step towards developing a coupled LDA system capable of exploiting current conventional SYNOP and future satellite observations.

The Harmonie-AROME model configuration used for our LDAS study focuses on using multi-layer surface physics and soil moisture analysis using the ISBA-DIF land surface model project. The code base used is from Harmonie-AROME CY46h1, incorporating multi-layer physics such as ISBA-DIF for soil processes using a 14-layer soil model. The upper-air data assimilation utilizes 3DVAR, while surface analysis includes EnSRKF (Ensemble Square Root Kalman Filter) LDA. The results are compared against analysis from sEKF LDA technique also.

- Experiment Details: Cold start on 1st June 2023 with 3-hour cycling for 8 weeks. Local settings for upper-air data assimilation are consistent across both LDAS runs.
- Surface Physics: The model employs ISBA-DIF for detailed 14-layer soil processes, including multi-energy balance (MEB) for vegetation effects.
- Kalman Filter Approaches: (a) EnSRKF for improved surface analysis. (b) sEKF for comparison.
- Observations used and other details: SYNOP, T2m and RH2m observations with defined errors (1K and 0.4%, respectively). The ensemble has 16 members, with a control run using deterministic forecasts. Control variables include soil temperature, TG1, TG2 (i.e. temperature at soil layer 1 and 2) and soil moisture, WG2 to WG5 (i.e. soil moisture variables from layers 2 to 5). Data assimilation is conducted every 3 hours. The experiments are performed over the “NORD_2.5km” domain, covering the Scandinavian peninsula (see Figure 1) .
- Experiments are also performed with an aim to evaluate the impact on the analysis quality of including more soil variables into the control vector. In this separate experiment the size of the control vector of soil moisture is increased with two additional variables, including soil moisture variables from layers 6 and 7 in addition to the reference configuration WG2 to WG5).
- Perturbation Methodology of Meteorological Forcing: Perturbations are applied following the methods described by Charrois et al. (2016) and Blyverket et al. (2019). Cross-correlated AR (1) process is used for generating perturbations. The fields perturbed include:

    a) Precipitation and shortwave radiation (multiplicative perturbations).

    b) Longwave downward radiation (additive perturbations).

    c) Soil moisture perturbations are also multiplicative, while soil temperature perturbations are additive.

The soil temperature and soil moisture differences in layer 1 (not shown), closest to the surface, are more pronounced. This suggests that meteorological forcing perturbations primarily impact the upper soil layers. As we move to deeper layers, the intensity of soil temperature and soil moisture differences diminishes. By layer 12 i.e. 5m depth in soil, there is little to no noticeable soil temperature and soil moisture difference, indicating that meteorological forcing perturbations have limited influence on deeper soil layers. The spatial variation of soil temperature and soil moisture differences show heterogeneous patterns . that could be due to regional climatic conditions, soil properties, or other factors influencing the response. The analysis of soil temperature and soil moisture perturbations shows that EnSRKF technique predominantly affects the temperature in upper soil layers, with limited impact at greater depths. The use of EnSRKF seems effective in capturing these spatial and depth-specific variations in soil temperature differences, which is valuable for improving soil temperature and soil moisture predictions in response to meteorological forcing.

Figure 1 compares the impact of meteorological forcing perturbations on latent heat flux and sensible heat flux in the NORD_2.5 km domain, using two different LDAS methods: the sEKF and the EnSRKF method. The sEKF and the EnSRKF data assimilation schemes are very different in their nature. The sEKF panels (labeled as "(a)" and "(c)") show minimal or negligible differences in both latent and sensible heat flux. This implies that the sEKF method has limited sensitivity to the perturbation of meteorological forcing by its construction when the infinitesimal perturbations are imposed on soil variables to numerically estimate Jacobians. The EnSRKF panels (labeled as "(b)" and "(d)") display substantial variations in both latent and sensible heat flux. This suggests that the EnSRKF technique is substantially more sensitive to perturbations of meteorological forcing and captures a broader range of spatial variations. The ensemble spread in the EnSRKF is directly related to the uncertainty measure in the soil analysis. The figure thus demonstrates that the EnSRKF method is far more sensitive to meteorological forcing perturbations than the sEKF, showing significant differences in both latent and sensible heat flux across the region. This makes EnSRKF potentially more suitable for capturing fine-scale variations in land-atmosphere energy exchanges.

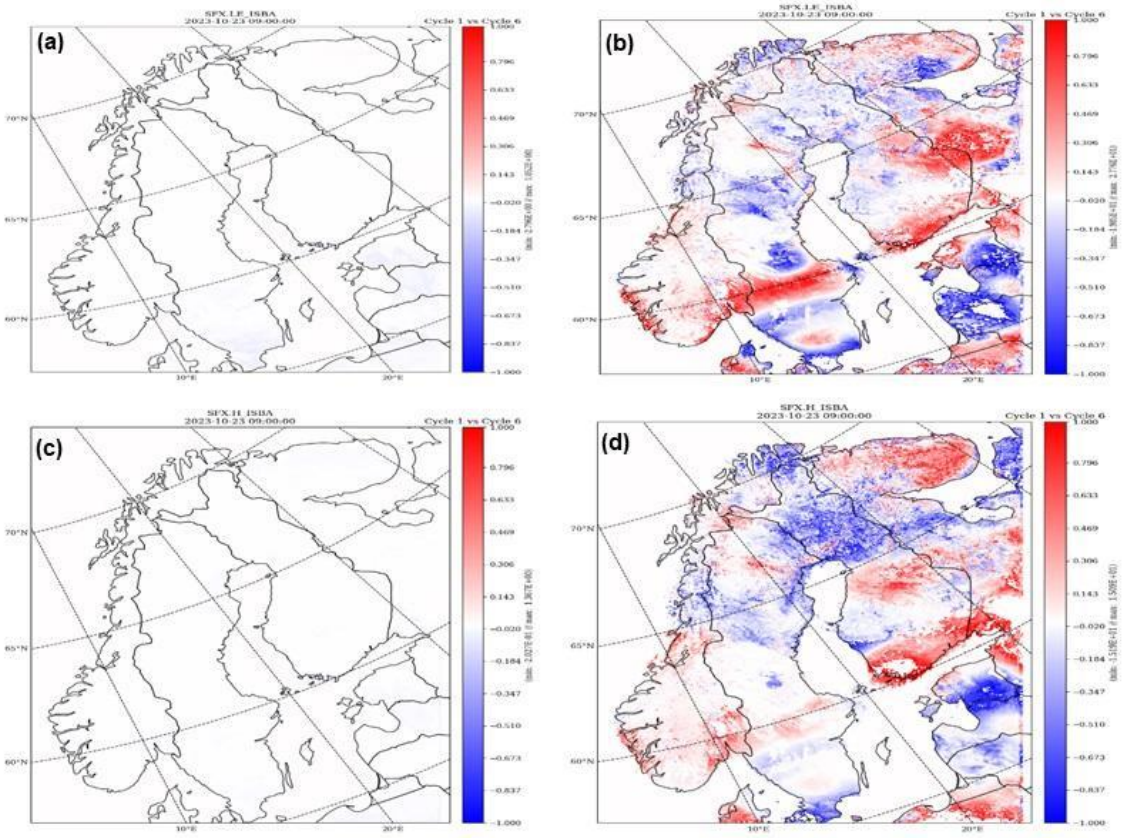

Figure 1: Perturbation of latent heat flux (W/m^2) over the NORD_2.5km domain produced by (a) sEKF (b) EnSRKF scheme, (c) and (d) same as in (a) and (b) but for sensible heat flux (W/m^2)

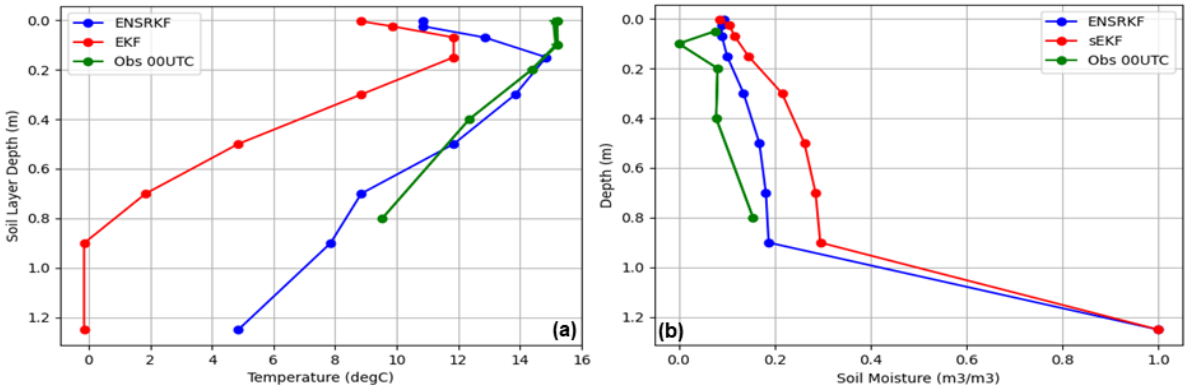

Figure 2: Vertical profile of (a) soil temperature and (b) soil moisture profile at Sodankylä station, LUO0009 from sEKF, EnSRKF based analysis and observations for a typical date of 00 UTC, 29062023.

Figure 2 presents the vertical profiles of soil temperature analysis (left panel, marked as (a)) and soil moisture analysis (right panel, marked as (b)) at the Sodankylä station, Finland with comparisons at 00 UTC of 29th June' 2023 of the EnSRKF (blue) and sEKF (red) and in-situ observations (green).

The temperature profiles show reasonable alignment between the observations (green) and the EnSRKF (blue) near the surface, especially in the top 0.2 m of soil. The sEKF (red) based profile**,** however, shows a notable difference, with consistently lower temperatures and higher soil moisture within this depth range. As the depth increases (between 0.4m – 0.8m), EnSRKF based temperature profile is matched with observations. Meanwhile, the sEKF has a generally colder profile and does not align with observations. The EnSRKF based soil temperature

profile better captures the overall gradient (the gradual increase of temperature with depth). In contrast, sEKF has a higher bias and does not capture the gradient as effectively, especially in the upper layers. In the top soil layer (up to 0.2 m), observations (green) show lower soil moisture values than both EnSRKF (blue) and sEKF (red). Both the sEKF based and the EsSRKF based analyses overestimate the surface soil moisture according to the independent observations, whereas the EnSRKF profile shows less pronounced overestimation. . Overall, the EnSRKF (blue) shows a closer match to observations than sEKF (red) for both soil temperature and moisture, especially near the surface. This suggests that EnSRKF might provide a more realistic representation of surface soil processes than sEKF. At the same time both EnSRKF and sEKF overestimate soil temperature inversion in the top layer. The reason for this is under investigation. One should also take into account that the point comparison between modelled and observed values are not straightforward because of the heterogeneity of the surface conditions not properly represented in the model.

The vertical profiles indicate that EnSRKF performs better than sEKF in replicating observed soil conditions, particularly near the surface. However, both methods show deviations from observations at greater depths, with EnSRKF showing a cold bias in temperature and both. models overestimating soil moisture in the upper layers. This indicates that EnSRKF is more suitable for near-surface applications, but further refinements of the scheme are needed to improve accuracy of the analysis at deeper levels. One should take into account that the memory of the soil variables at deeper soil levels is relatively long and two weeks period used for spinning up of structures might be not sufficiently long for the analysis of the soil variables at the deep layers. The deepest soil layers are not so sensitive to meteorological conditions and thus are not updated during the analysis.

For the near surface variables dew-point temperature, specific humidity, and 2-meter air temperature, the EnSRKF LDAS based forecasts show a noticeable reduction in bias compared to the sEKF approach. The standard deviation of the forecasts is also lower for EnSRKF LDAS based forecasts, indicating a more accurate prediction. This reduction in bias and improved consistency in forecast accuracy is valid from 3 up to 30 hours forecast length. This emphasizes the effectiveness of using an ensemble-based approach for LDAS in improving forecasts of key near-surface meteorological variables.

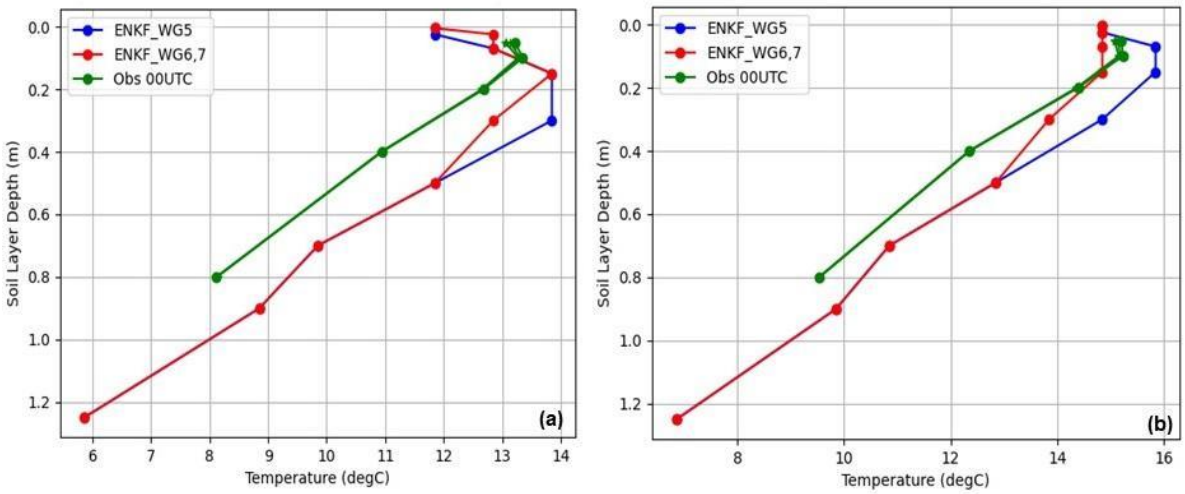

Figure 3: Vertical profile of soil temperature profile at Sodankylä station, LUO0009 from EnSRKF LDAS experiments, ENKF_WG5 (control vector TG1, TG2, WG1 to 5) and ENKF_WG6,7 (control vector TG1, TG2, WG1 to 7) based analysis and observations for two typical dates (a) 00 UTC, 23062023 (b) 00 UTC, 29062023.

Figure 3 shows the results from experiments using EnSRKF based LDAS with two configurations, ENKF_WG5 (control vector TG1, TG2, WG2 to 5) and ENKF_WG6,7 (control vector TG1, TG2, WG2 to 7) show that the configuration ENKF_WG6,7 which assimilates observations up to WG7 soil moisture layer, shows better alignment with observed temperatures, especially near soil layers closer to surface. Thus, enlarging the dimensionality of LDAS control vector results in a closer match with observed temperature profiles, highlighting the importance of including additional deeper soil layers in the data assimilation process. The mechanism behind this is under investigation. One possible reason is the larger dimensionality of the control vector that allows better capture memory of the nonlinear system.

Figure 4 shows a comparison of soil temperature analyses produced by the EnSRKF and sEKF LDAS against independent observations at 5 soil depth layers at Sodankylä station (LUO0009). Each subplot represent time series of soil temperature at specific depths: 5cm, 10 cm, 20cm, 40 cm and 80 cm.

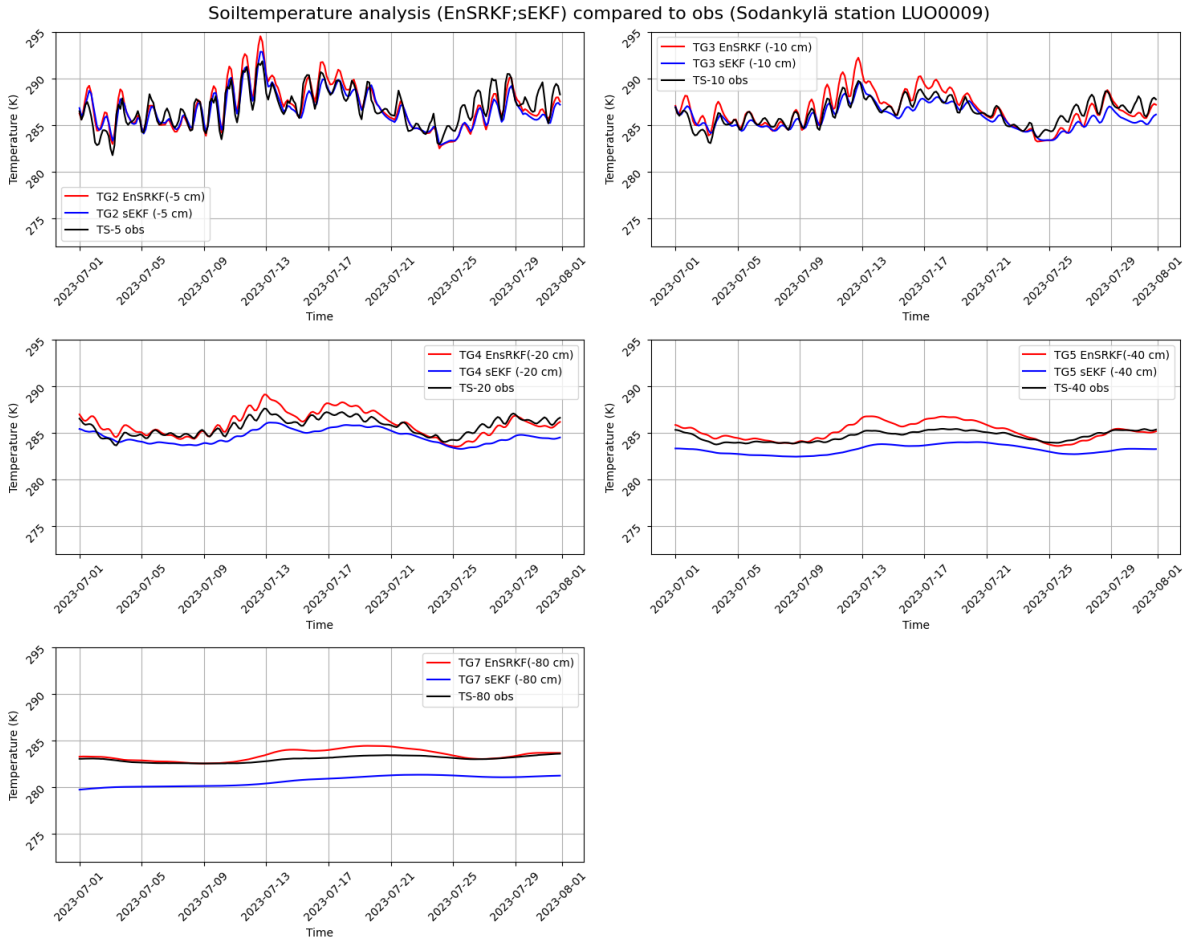

Figure 4: Soil Temperature Analysis provide by the EnSRKF (red) and the SEKF (blue) LDAS against independent observations (black)

At 5cm depth EnSRKF has slightly more variability and overestimates the temperature at some peaks, while sEKF aligns more closely with observations in certain instances. At 10 cm. the EnSRKF aligns better with observed peaks while the sEKFis slightly over-smoothed, suggesting the strength of the EnSRKF in representing rapid changes in soil temperatures. At 20 cm depth both the systems show dampened fluctuations compared to shallower depths. In general, the EnSRKF system is more effective at capturing temperature dynamics in the middle soil layer. In the soil temperature profile at 40 cm depth, both EnSRKF and sEKF maintain consistency, with minimal differences between their modeled values. Whereas at 80

cm depth the EnSRKF based soil temperature time series is closer to observations reflecting stable deep soil temperature dynamics. This analysis time series of soil temperature highlights the performance of EnSRKF and sEKF in modeling soil temperature across layers. The sEKF based analysis is less efficient in propagating the information from the screen level variables to the deeper soil levels (40 cm and 80 cm) and underestimate soil temperature, while the EnSKRF analysis is too sensitive to the fluctuations of screen level variables and tend to overestimate soil temperature at the deeper levels. Note that the experiment is performed during summer period and meteorological conditions contribute to the warming.

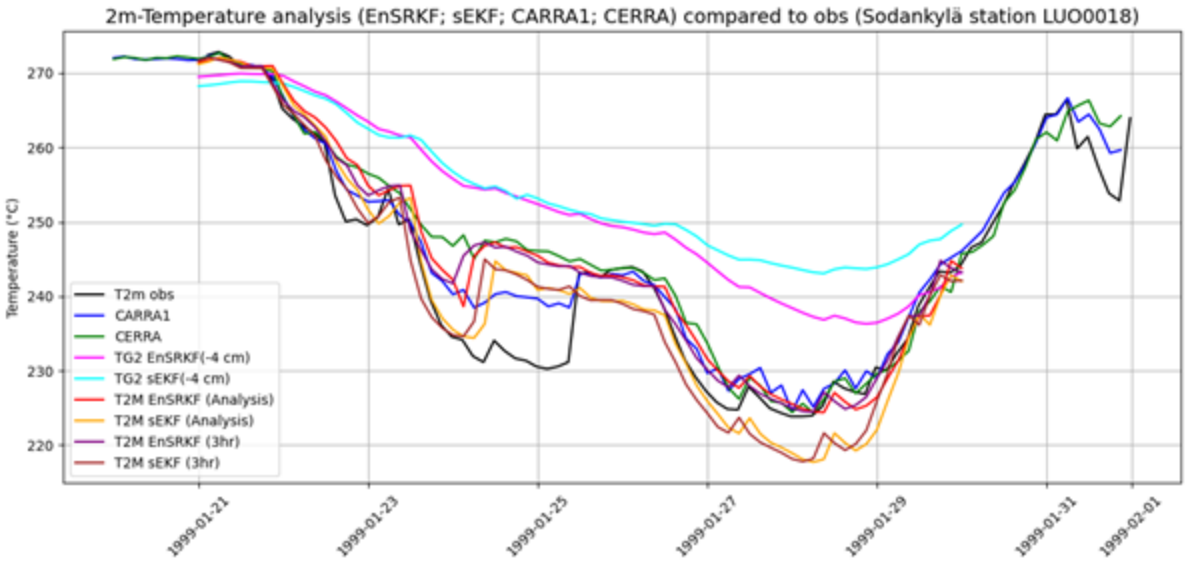

Figure 5: Time series of 2-meter air temperature (T_{2M}) from observations, reanalysis systems (CARRA1 and CERRA), and analysis/forecasts (3hr forecast) from EnSRKF and sEKF based LDAS, soil temperature (TG2) at 4 cm depth at Sodankylä station (LUO0018) during January 1999.

The figure 5 represents a time series comparison of 2-meter temperature analysis and observations at Sodankylä station (LUO0018) during January 1999 (winter period). The figure highlights the performance of different LDAS modeling approaches in capturing temperature variations during a cold spell. T_{2M} observations (black line) represents ground truth data recorded at the station. It is used as a reference for comparison. During most of the occasions of the significant drop of temperature (January 22,23,24,27) all systems capture this trend well, but with some variations in magnitude and timing. The CARRA1 and CERRA reanalysis results are from OI based LDAS and show smoother curves compared to the more variable EnSRKF and sEKF based land analysis and forecasts. The soil temperature analysis from EnSRKF and sEKF based LDAS shows initial trend closer to observations, especially during extreme cold periods reflecting deep inversion up to 230K. Both EnSRKF and sEKF analysis (red and orange lines) capture the temperature trends well, reflecting their ability to integrate observations effectively. The T_{2m} from EnSRKF (Analysis) shows good agreement with observed trends, capturing the main temperature changes and extremes. However, minor deviations may indicate limitations in the assimilation process or differences in how surface conditions are modeled. The deviation between sEKF and EnSRKF analysis reflects differences in how the two systems assimilate data and adjust the model's state. Similarly, the 3-hour forecasts show slight deviations from their respective analyses but generally maintain the same trend, indicating good short-term predictive skill. The comparison suggests that both EnSRKF and sEKF systems have strengths in short-term forecasting, but their performance varies slightly depending on the assimilation methodology and model configurations.

We have noticed that the EnSRKF provides a somewhat stronger response to screen level observations in particular in connections to the daily cycle than the observations suggest. This

results in too large increments of control variables in the deeper soil layers. More diagnostics and investigations are needed to identify the reason for this behavior. Several experiments are being conducted with the aim to investigate possible remedies to this behavior both through tuning of the filter characteristics such as observation error standard deviation and ensemble spread and the reducing/eliminating origin of systematic errors. The evaluation of the model simulations against soil moisture observations are in particular challenging because of inconsistencies between predicted precipitation patterns and those observed.

Work is also ongoing in introducing snow control vectors in the multi-layer snow scheme initialisation within the inline EnSRKF based LDAS in the HARMONIE-AROME system.

3.1.2 Local Ensemble Transform Kalman Filter in the stand-alone offline environment.

The local ensemble transform Kalman filter (LETKF) is an attractive approach for land data assimilation being relatively easy to implement and flexible for choice of observation and control variables. It is parallelizable and offers flow dependent background error structures.

At each grid point, the filter equations are solved in ensemble space and the analysis increments are a linear combination of the background ensemble perturbations. In this work we follow the implementation in (Hunt et al., 2007).

$$\begin{aligned}
\mathbf{x}^a &= \bar{\mathbf{x}}^b + \mathbf{X}^b\mathbf{w}^a \\
\mathbf{w}^a &= \mathbf{W}^a + \bar{\mathbf{w}}^a \\
\bar{\mathbf{w}}^a &= \tilde{\mathbf{P}}^a(\mathbf{Y}^b)^T\mathbf{R}^{-1}(\mathbf{y}^o - \bar{\mathbf{y}}^b) \\
\mathbf{W}^a &= [(k-1)\tilde{\mathbf{P}}^a]^{1/2} \\
\tilde{\mathbf{P}}^a &= [(k-1)\mathbf{I}/\rho + (\alpha \circ (\mathbf{Y}^b)^T\mathbf{R}^{-1})\mathbf{Y}^b]^{-1}
\end{aligned}$$

where x represent the ensemble control vector, X the ensemble anomalies, *a* and *b* superscripts indicate analysis and background respectively, w is the transformation weights between the background and the analysis, Y^brepresent the ensemble observation equivalent. $P$and R are the error covariance matrices. ρ and α are tunable parameters for inflating the background error covariance matrix and to apply localization (inflation of R), respectively. Figure 6 gives an example of the W^a transformation matrix.

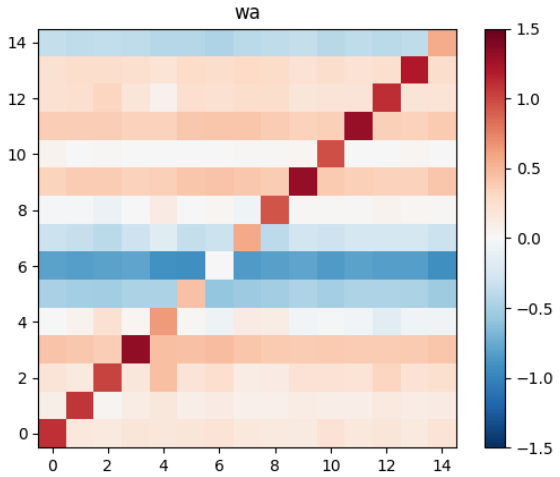

Figure 6: Example of the W^a transformation matrix with dimensions n times n, where n is the ensemble size.

A prototype of the LETKF is implemented in a python package (https://github.com/CERISE-Regional-Dem/sfcpert). The package also includes a number of pre and post processing tools which are required to perform the soil and snow analysis for the SURFEX surface model. However, an effort is made to keep the filter independent and general for possible other usages. There is ongoing activity to develop the code for better scalability, and efficient use of resources. One of the attractive features of the LETKF scheme is flexibility modelling footprint operator for future assimilation of satellite observations into SURFEX model. Experiments presented here are conducted using screen level observations.

Localization

The LETKF indirectly uses the ensemble correlation between state variables in a grid point and the observation equivalent to spread information spatially. Having a finite ensemble size with limited representation of the true spatial structures, a localization is required to avoid sampling noise. For the localization parameter α, we use an exponential decay function with an e-folding length of 50 km. Since e.g. snow depth, temperature and humidity observations can have limited representativeness at other altitudes, the vertical distances should also be taken into account. In our implementation, the same exponential decay is used for vertical localization as horizontal. But then with an e-folding length of 200 meters. LETKF method is tested for data assimilation into the multi-layer snow model.

Ensemble generation

Special treatment of the snow pack and bounded variables

To achieve optimal results for a snow analysis a number of pre and post processing steps are implemented. The snow model (ISBA-ES) represents the snow pack state with snow water equivalent (swe), density (rho), heat, age, and albedo. For gridpoints with no snow, swe is equal to zero, and all other variables undefined. In order to produce an analysis, it is thus necessary to initialize the undefined variables to realistic values. A choice was made to use the ensemble mean where this exists. Using the ensemble mean value ensures that no extreme increments enter the system. In the case where all members have no snow, the filter will not be able to produce other analysis than zero. If the latter turns out to be a weakness,

the ensemble perturbations need to be reconsidered to capture the probability of snow. Or, other more pragmatic solutions can be considered.

Most soil and snow variables have a physical bound to their values. After the LETKF is performed, all relevant variables are checked against predefined limits. Members that violate some limit are flagged, and then replaced by the ensemble mean of the remaining healthy members.

Verification experiment

Assimilation of point observations involves challenges related to representativeness of observations which are difficult to distinguish from background error structures. In order to minimize this uncertainty, and to have a truth to compare with, synthetic experiments were set up. A reference experiment was run using forcing data from a different source to act as a truth and to provide synthetic observations to the data assimilation experiment. Point observations were simulated by interpolating model values from the reference run to real observation locations. These observations were further assimilated into the data assimilation experiment. This experimental setup allows for validation in model space and assesses the performance away from observation locations.

Table 1: Tested configuration of unified system

Control Variables	Observations	Perturbations
snow state (12 levels) - swe - rho - heat soil state - water content (5 levels) - temperature - (3 levels)	- synop 2m temperature - in situ snow depth	- Rain fall - Snowfall - temperature - shortwave radiation - longwave radiation

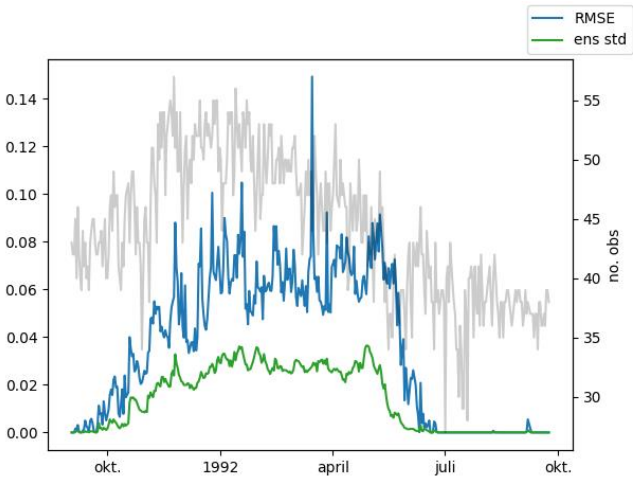

Figure 7. Spread - skill relationship for Snow depth in the real observation experiment. Grey line indicates the number of observations.

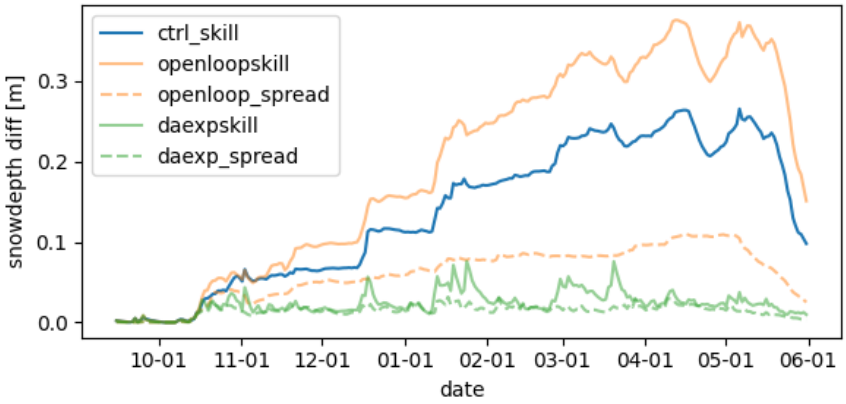

Figure 8. Spread - skill relationship for snow depth in the synthetic experiment (From report D1.1)

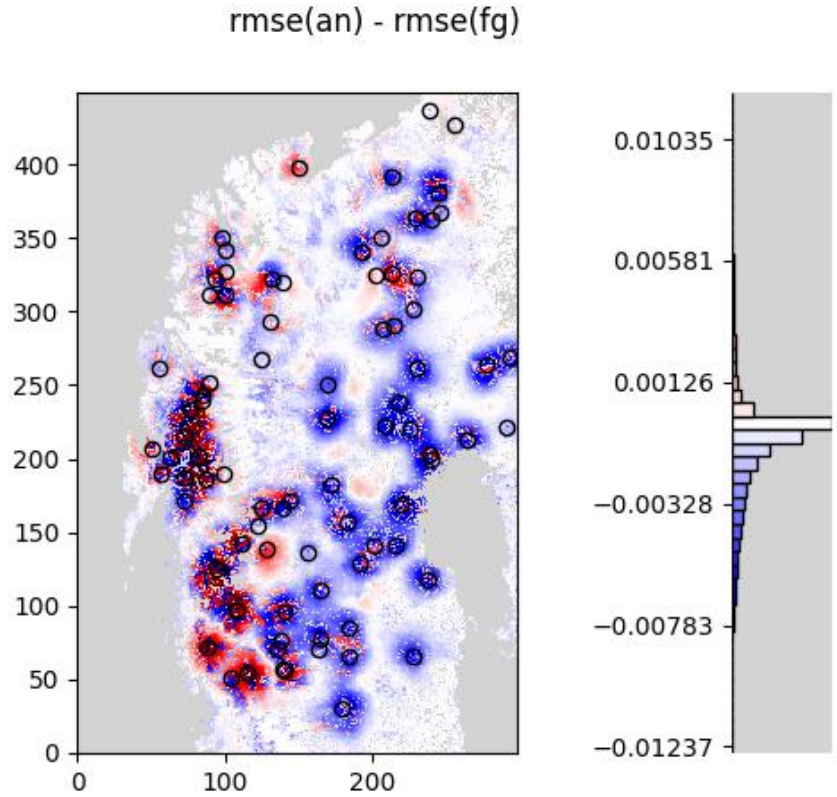

Figure 9. RMSE difference indicating whether increments improved or degraded the state estimate

In observation space the filter is able to correct snow depth and follow the observations closer than the open loop during a full snow season Figure 8 . In general the ensemble encapsulates the observations. However, for some stations, rapid changes in observations are not captured by the analysis, which could indicate that the forcing perturbations were not large enough. 2 meter temperature perturbations were added as a measure for this behavior. The relationship between model errors (skill) and ensemble spread was also assessed. In the synthetic experiment, where representativeness errors can be disregarded, the spread - skill relationship was close to unit, and very satisfactory (Figure 8). However, in the real observation experiment, the results indicate an under-dispersive ensemble (Figure 7). This is likely caused by the representativeness errors of the observations, and also that the forcing

data is further away from the real weather than the reference simulation used for the synthetic observations.

Spatially, the system is able to improve the state in an area surrounding the observation point within the localization length, particularly in inland areas with flat topography (Figure 9). Areas where the system fails to improve the state are especially found in regions with mountains and valleys. This indicates that the spatial structure of the ensemble perturbations are not sufficiently realistic. The issue over mountainous regions can be approached by tuning ensemble perturbation and tuning of horizontal and vertical localization length scales.

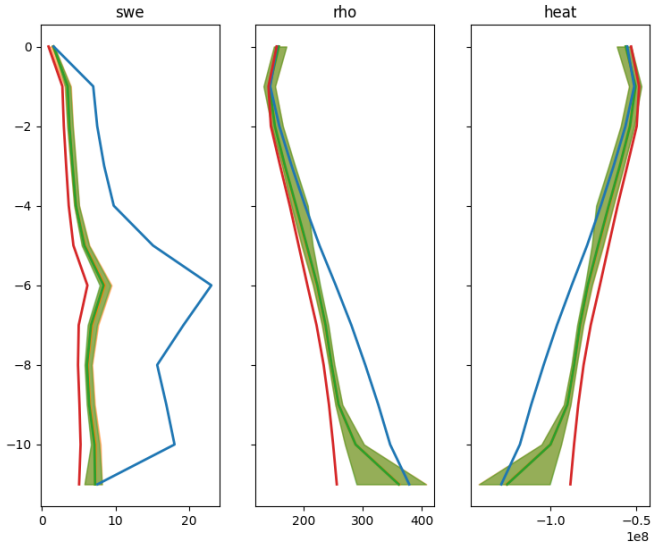

Figure 10: Snowpack profiles demonstrating improvement in unobserved state variables. Truth (red), open loop (blue), analysis (green)

Profiles of the snow pack are also compared at a selection of observation points (Figure 10). A majority of these profiles showed analyses closer to the truth than first-guess, however, some showed degradations. Since the ensemble perturbations in snow state are results of the perturbed forcing, it is critical that the ensemble forcing is close and includes the true weather conditions. This is most likely not always the case, given the limited number of ensemble members and the assumptions behind perturbation generation. Snow and soil are challenging domains due to the long memory of past weather and thus past perturbations.

The synthetic experiments demonstrated that the LETKF is suited for the purpose of land surface data assimilation and showed overall satisfactory results to continue with the development. Localization parameters were adjusted based on the experiment to limit the horizontal impact of each observation.

LETKF surface data assimilation in an inline setup

The ensemble-based land surface data assimilation (DA) systems described in this report are run offline (no coupling to the atmosphere) to create the land surface ensemble spread (even in the inline framework). From WP1 D1.1 we found that for the assimilation of screen level variables this offline ensemble introduces spurious correlations that could degrade the land surface analysis. In particular we saw that the ensemble correlation between soil moisture (layer 1) and 2m specific humidity was negatively correlated, while for a coupled ensemble (EPS) this correlation was positive, see Figure 11.
A consequence of this could be that if the observation-minus-forecast of humidity is larger than zero (model too dry), the land DA could remove soil moisture and add it to the atmosphere (which is not coupled), potentially drying the surface. At the same time deeper investigations need to be performed to understand the reason for this behaviour. The relation between soil moisture and T2m and Q2m is complex and is prone to memory/phase delay. As experiments done in Section 3.1.1 show the dimensionality of the control vector has an impact on the ability of the Ensemble Kalman Filter to capture memory of the nonlinear system.

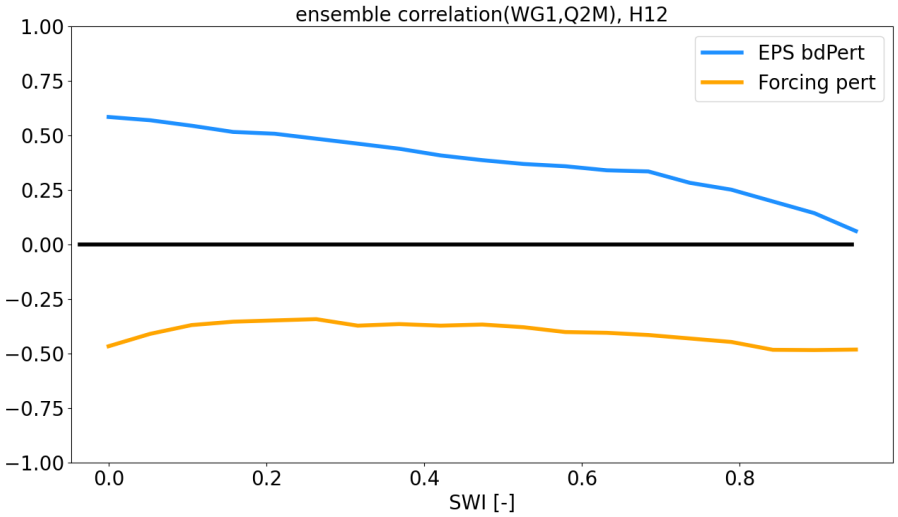

Figure 11: Ensemble correlation between soil moisture (layer 1) and 2m specific humidity for different soil wetness index (SWI). For the coupled ensemble system (EPS bdPert in blue) and offline perturbed forcing experiment (Forcing pert in orange).

Focusing on coupled land-atmosphere DA, it is also relevant to investigate how the coupled ensemble represents the covariance between variables, both in the land surface and between surface and upper-air variables. In this section we report preliminary results on assimilating screen level variables with an offline vs an inline ensemble.

Experiment setup and results

For the surface analysis part we use the LETKF scheme. It is slightly adopted to read and write FA format files which is the internal file format in HARMONIE-AROME. Other than that no specific modifications were done to the scheme.
We set up two different experiments, i) with a 7 member inline EPS (driven by different ECMWF boundary conditions) and cycling of the surface, ii) 16 member offline surface

ensemble, cycling of surface and with perturbation of soil temperature and moisture. The experiments are run from the 1st July 2023 to the 3rd July 2023.

Figure 12 shows the ensemble spread in soil temperature for layer 2 for inline vs offline ensemble, respectively. From the figure 11 we can see that the spread in the inline ensemble increases throughout the period. While for the offline the spread is relatively flat throughout the period. In Figure 12 we plot the same time-series but for soil moisture layer 2. Here we see that there are larger differences between the two ensemble methods, in particular when it comes to timing of the precipitation events.

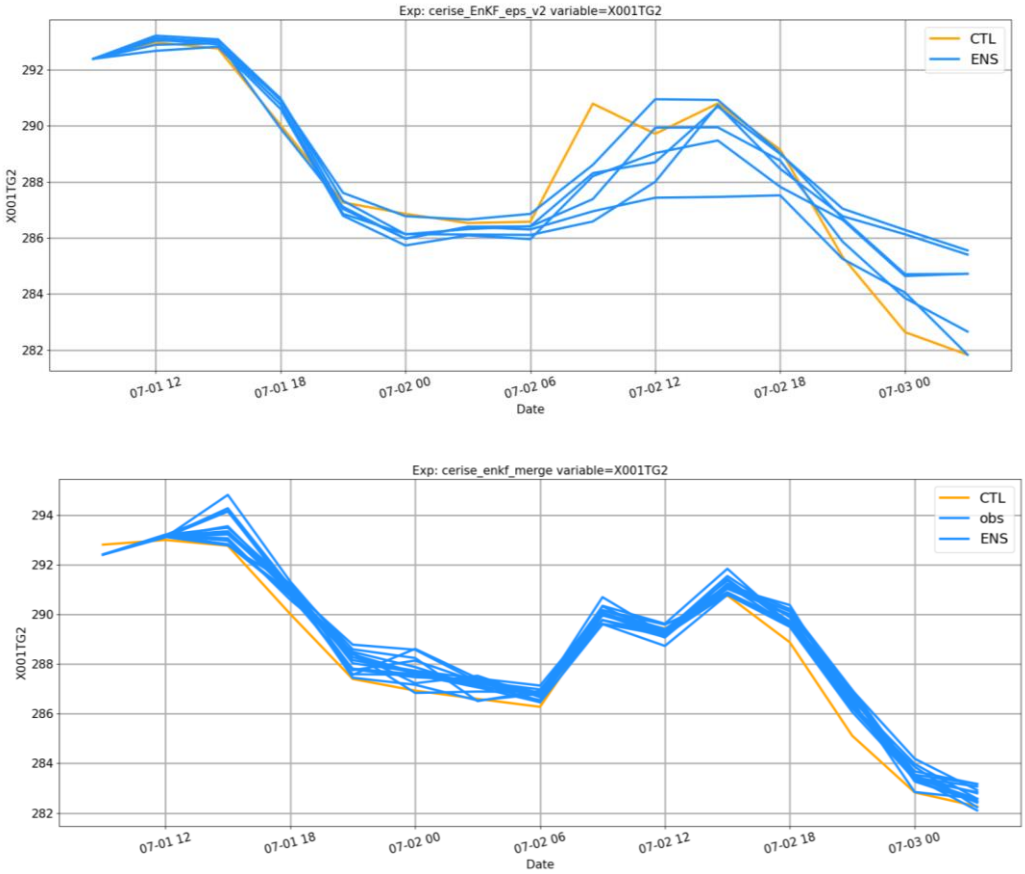

Figure 12: (Top) soil temperature (layer 2) spread inline ensemble, (bottom) same but for offline ensemble. Experiment period is 1st July 2023 06UTC to 3rd July 2023 06 UTC

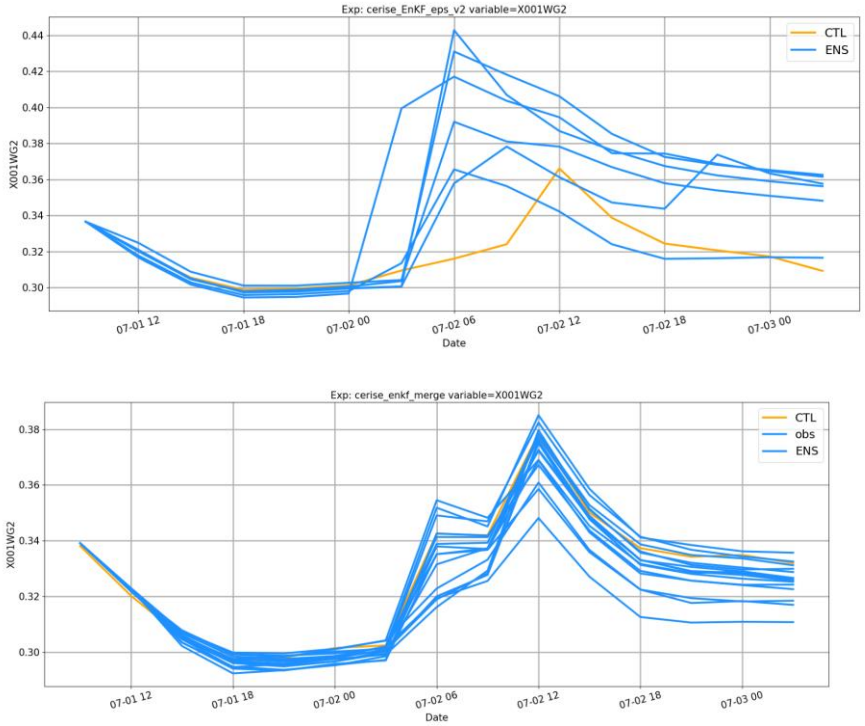

Figure 13: (Top) soil moisture (layer 2) spread inline ensemble, (bottom) same but for offline ensemble. Experiment period the 1st July 2023 06 UTC to the 3rd July 2023 06 UTC

The resulting spread in the screen level variables are shown in Figure 12 and Figure 13. It is clear that the offline ensemble is under-dispersive and that the coupled ensemble spread is larger and encapsulates the observed value.

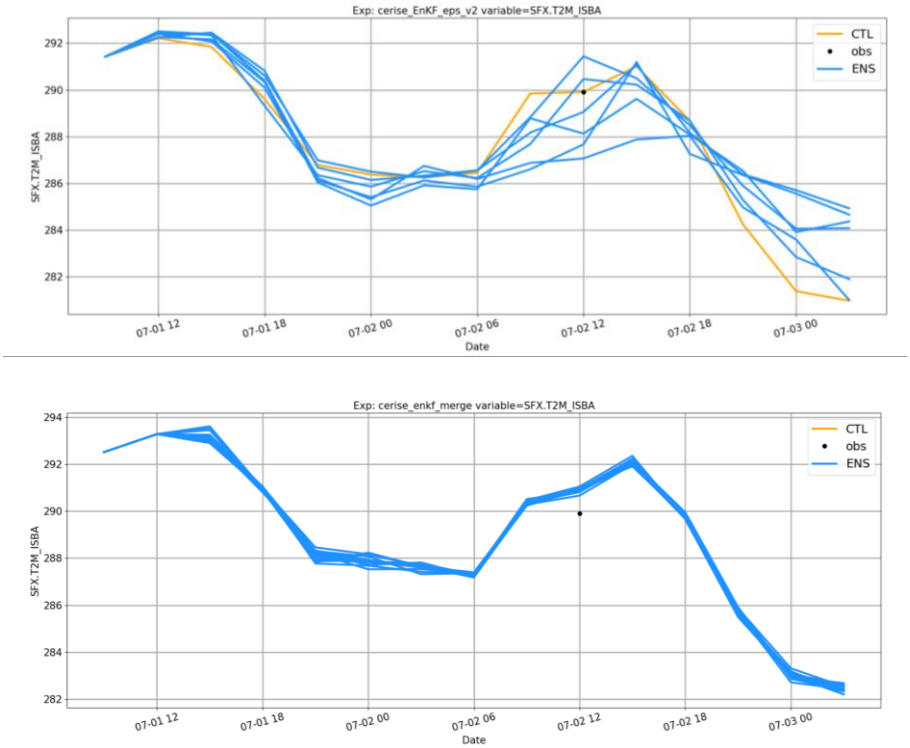

Figure 14: (Top) 2m temperature spread inline ensemble, (bottom) same but for offline ensemble. Observed value given by black dot. Experiment period 1st July 2023 06 UTC to the 3rd July 2023 06 UTC

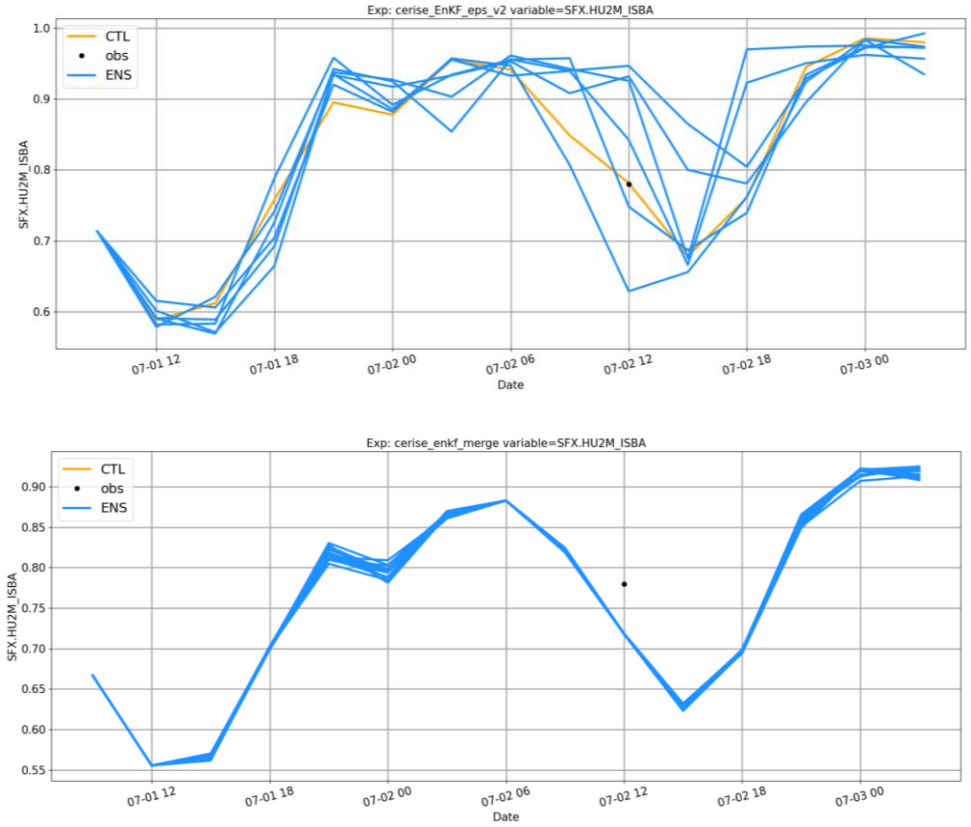

Figure 15: (Top) 2m relative humidity spread inline ensemble, (bottom) same but for offline ensemble. Observed value given by black dot. Experiment period the 1st July 2023 06 UTC to the 3rd July 2023 06 UTC

We performed single observation experiments to evaluate how the coupled ensemble would affect the land surface analysis. The EPS ensemble was then used as the first guess to the LETKF and we assimilated 2m temperature and relative humidity. The control vector was soil temperature layer 1-2 (TG1 and TG2) and soil moisture layer 2-5 (WG2-WG5). The observation-minus-forecast difference was 0.65 K and -0.03 for temperature and humidity, respectively.

Figure 16 shows the soil temperature and moisture increments for different layers. The soil temperature increment is quite straightforward, positive Observation-minus-first-guess (OmF) difference and positive increment which decays into the soil. For soil moisture the increments are more complex and layer 3 seems to have larger increments than layer 2 for some regions. This is most likely related to the rooting depth and hence which layer is active in the transpiration. The soil moisture increments are quite large along the coast.

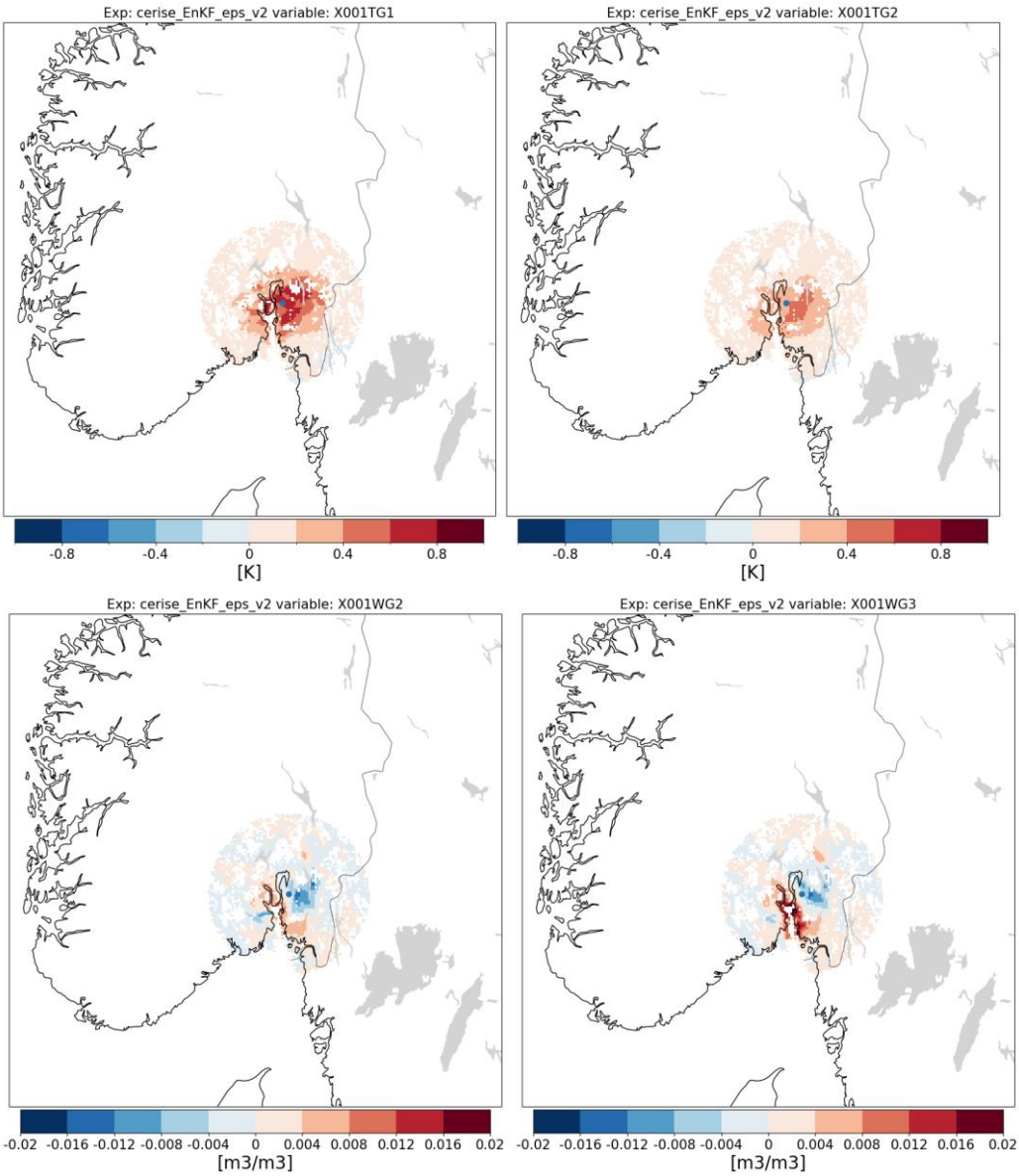

Figure 16: (Top) Soil temperature increments layer 1 and 2, (bottom) soil moisture increments layer 2 and 3.

This might be a spurious increment and would need to be investigated in more detail later.

To further evaluate the sanity of the ensemble covariance structure. We compute the Kalman gain:

$$K = Cxy/(Cyy + Cvv)$$

Here Cxy is the sample cross-covariance between the ensemble of prior model states and the predicted measurements, Cyy is the sample covariance of the predicted measurements and Cvv is the measurement error variance. The latter is set to $1K^2$ and 0.16 for 2m temperature and 2m humidity, respectively. In Figure 17 we show the Kalman gain for soil moisture and 2m temperature (left) and soil moisture and 2m humidity (right).

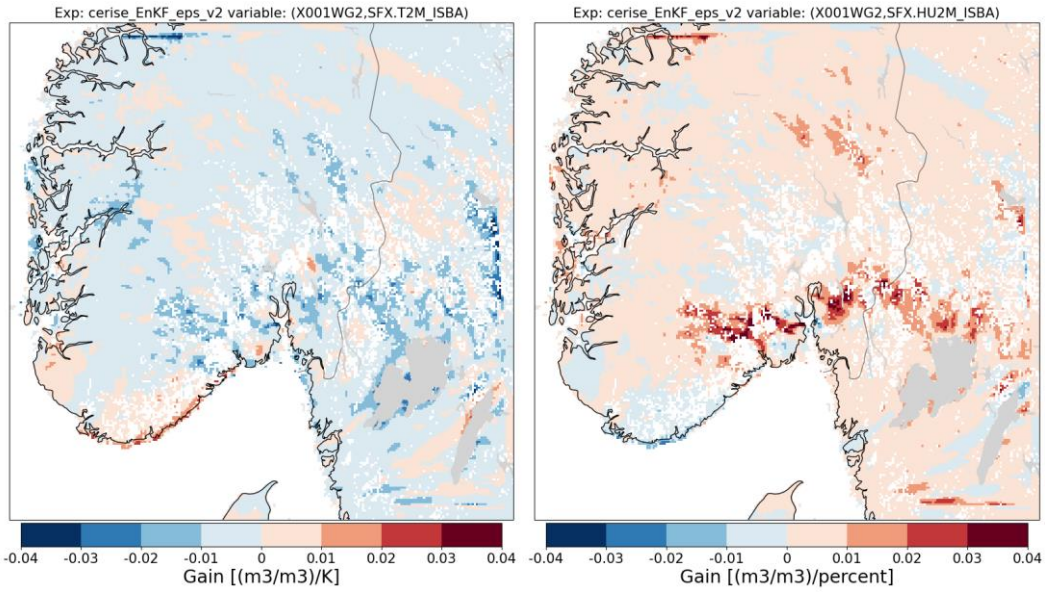

Figure 17: (Left) Kalman gain for soil moisture and 2m temperature, (right)) same but for soil moisture and 2m humidity.

For temperature a gain of K = -0.02 (m3/m3)/K, equates to a decrease of 0.02 (m3/m3) in the updated (posterior) layer 2 soil moisture for a difference of 1 K in the modeled vs observed 2m temperature. This seems reasonable to remove soil moisture to increase the sensible heat flux (and 2m temperature) and vice versa. For humidity we see both more of positive and negative gain, but the most dominant is the positive gain. For humidity a gain of K = 0.02 (m3/m3)/percent, equates to an increase of soil moisture for a positive difference in the observed vs modeled humidity. This also seems reasonable as a positive OmF difference means that the model is too dry and we add soil moisture available for evaporation.

3.1.3 Autoencoder based data assimilation in a test environment

One of the challenges in ensemble data assimilation is to generate realistic and diverse model state perturbations that capture the uncertainty in the initial conditions. Multivariate consistency of model state perturbations is another topic that was investigated in the study. The ultimate goal of this study is to propose a system that would be generic enough to handle a variety of situations in an unified environment. As it was already mentioned in section 3.1.1 and section 3.1.2 the dimensionality of the control vector impacts the ability of the DA system to handle memory. This means that the perturbations should respect the physical and dynamical relationships among different variables in the model state and this is situation-dependent. Sampling in the Latent Space using a Variational Auto-encoder (VAE) is one of the approaches that we have been investigating for this purpose. A VAE is a type of neural network that can learn a low-dimensional representation of high-dimensional data, such as images, texts, or model states. It consists of two parts: an encoder and a decoder. An encoder allows to encode a large dimensional physical space into a well-behaving smaller-dimensional latent space where sampling can be performed in an easy way. The decoder allows to project large size ensembles in an efficient way back to a physical space. In our study, we applied the VAE framework to ten years of data from a single column (located in Sodankylä in northern Finland) of the surface model SURFEX. Here SURFEX was run with a 14 layer diffusion scheme for the soil (ISBA-dif) and a 12 layer representation of the snowpack (Explicit snow) over two surface patches (high and low vegetation). In total the model state then contains about 200 variables.

Using principal component analysis, approximately 99 % of the variance could be explained using 32 variables. A vanilla VAE was shown to do the same with a latent space of only 10

variables. However, the time evolution of the reconstructed data becomes notably smoothed. This is a known problem with standard VAE and future work will be directed towards combating this, either by employing the Wasserstein distance or by using a deterministic auto-encoder (less prone to smoothing) followed by a step with diffusion or normalizing flow to guarantee compactness of the representation. One of the aims of this study was to obtain latent space that could capture both soil and snow processes. Up to now the results are not satisfactory. The reason is that when snow is present the snowpack heavily isolates the soil, and the correlations between screen level observations and the soil variables are negligible in presence of snow. Present solution is to use two different models for updating soil variables when there is no snow and for updating snow variables.

3.2 Ensemble Kalman Filter for hydrological model HYPE

In this section we present development of the Ensemble Kalman filter for the hydrological applications. The ultimate goal is the development of the consistent meteorological-hydrological data assimilation necessary to properly capture the water cycle evolution including snow. The Hydrological Predictions for the Environment (HYPE) model is a semi-distributed catchment-based hydrological model, developed and used for research and operational forecasting/analyses by SMHI hydrology group (Lindström et al. 2010). A module for data assimilation using the Ensemble Kalman Filter (Evensen 2009) has been previously implemented in the HYPE model as described in (Musuuza et al. 2020; Musuuza et al. 2023). The current EnKF implementation in the HYPE model is characterized by:

- ensemble generation is obtained by random perturbation on the meteorological forcing data (precipitation, air temperature, wind speed/direction) assuming normally distributed errors with either relative or absolute standard deviations. The perturbations can further be generated with temporal memory and with spatial correlation for individual variables, but without covariance between variables.
- observation ensembles are generated without spatial correlation, but otherwise with the same error model assumptions as for the forcing data.
- localization is obtained using horizontal and vertical distance-based covariance localization.
- the EnKF analysis is performed globally using all available observations and all spatial compartments in the model domain which limits the application to smaller model domains due to numerical demands.

The objective of this task is to further improve the EnKF implementation in the HYPE model through the following steps:

- implementation of a local EnKF to enable applications with larger model domains (inspired by LETKF developments for meteorological application)
- development and implementation of an improved hydrologically constrained localization method for assimilation of instream observations such as river discharge and river water level
- diagnose the sensitivity of EnKF scheme to a range of tunable parameters, tune the EnKF scheme on optimal performance and get better insight into the properties of Ensemble Kalman Filtering.

A challenge for hydrological data assimilation using the EnKF method is to properly take into account the spatio-temporal relations between different types of model states and observations. This is particularly the case for the assimilation of river discharge or river water level observations, which represent the integrated effect of hydrological processes in the

upstream area of the observation locations that may span various periods back in time depending on the basin size and memory characteristics. Based on (El Gharamti et al. 2021) we will implement and evaluate the impact of an along-stream localization method taking into account the relative positions of the observations and the model states along the river network rather than the horizontal and vertical distance used in the current localization method. In addition to (El Gharamti et al. 2021) we will include the relative control of the upstream area of a model state versus an instream observation, as well as the travel time along the stream from the location of an upstream model state to the location of the observation. Furthermore, we will investigate the potential of the empirically based localization method presented by (Revel et al. 2019) to further improve the assimilation of river discharge and water level observations.

3.2.1 Study areas

The first study area used is the Lake Överuman catchment, a small mountain catchment in Northern Sweden used for hydropower (Clemenzi et al. 2023). In this area, observations of snow depth and snow water equivalent are available with high spatial resolution along so-called snow survey transects, in addition to the lake (reservoir) inflow. Figure 18 shows the reservoir and the inflow region together with buffer area.

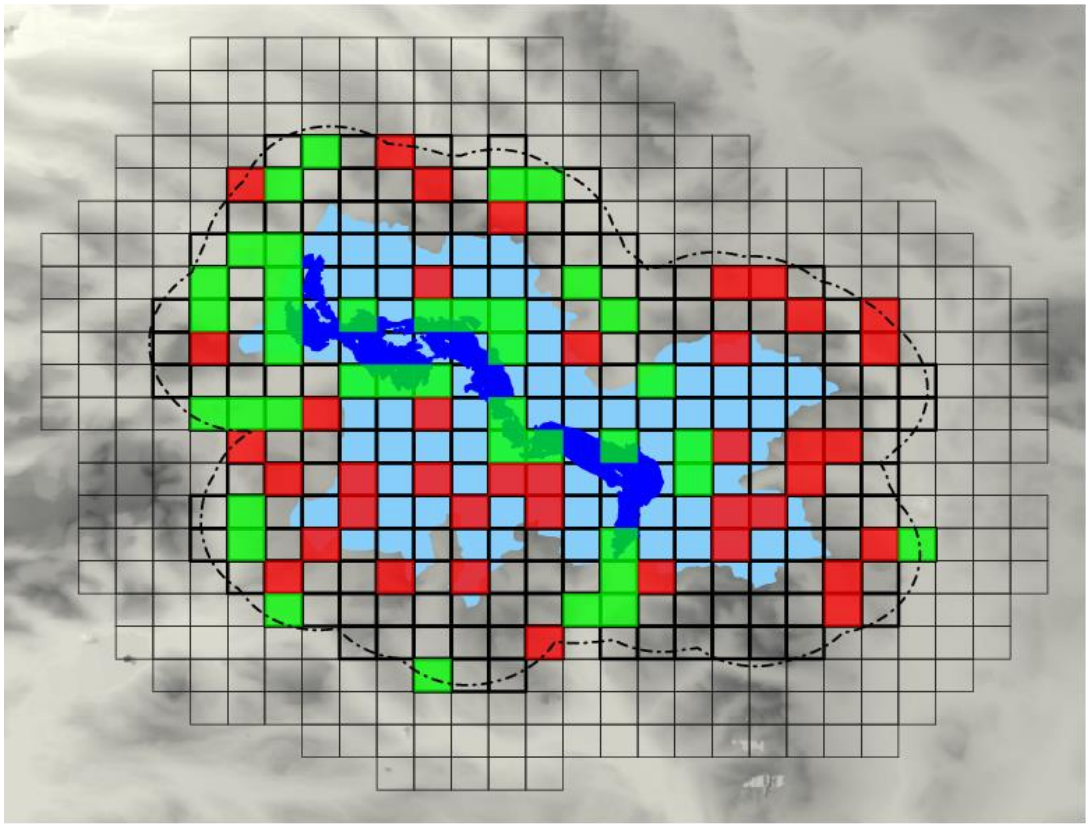

Figure 18: The study area used in the experiment marked as dark squares. The reservoir (dark blue) is shown with its inflow region (light blue) and the 6 km buffer (dashed line) around it. The green and red squares are the sub basin sampled from the bands below and above the tree line, respectively.

A second study area has been created by setting up the HYPE model on a 0.05-degree grid covering a larger part of northern Sweden. For this study area, see Figure 19, we will use river discharge and snow depth observations from the SMHI station observation networks, and fractional snow cover from the ESA CCI project.

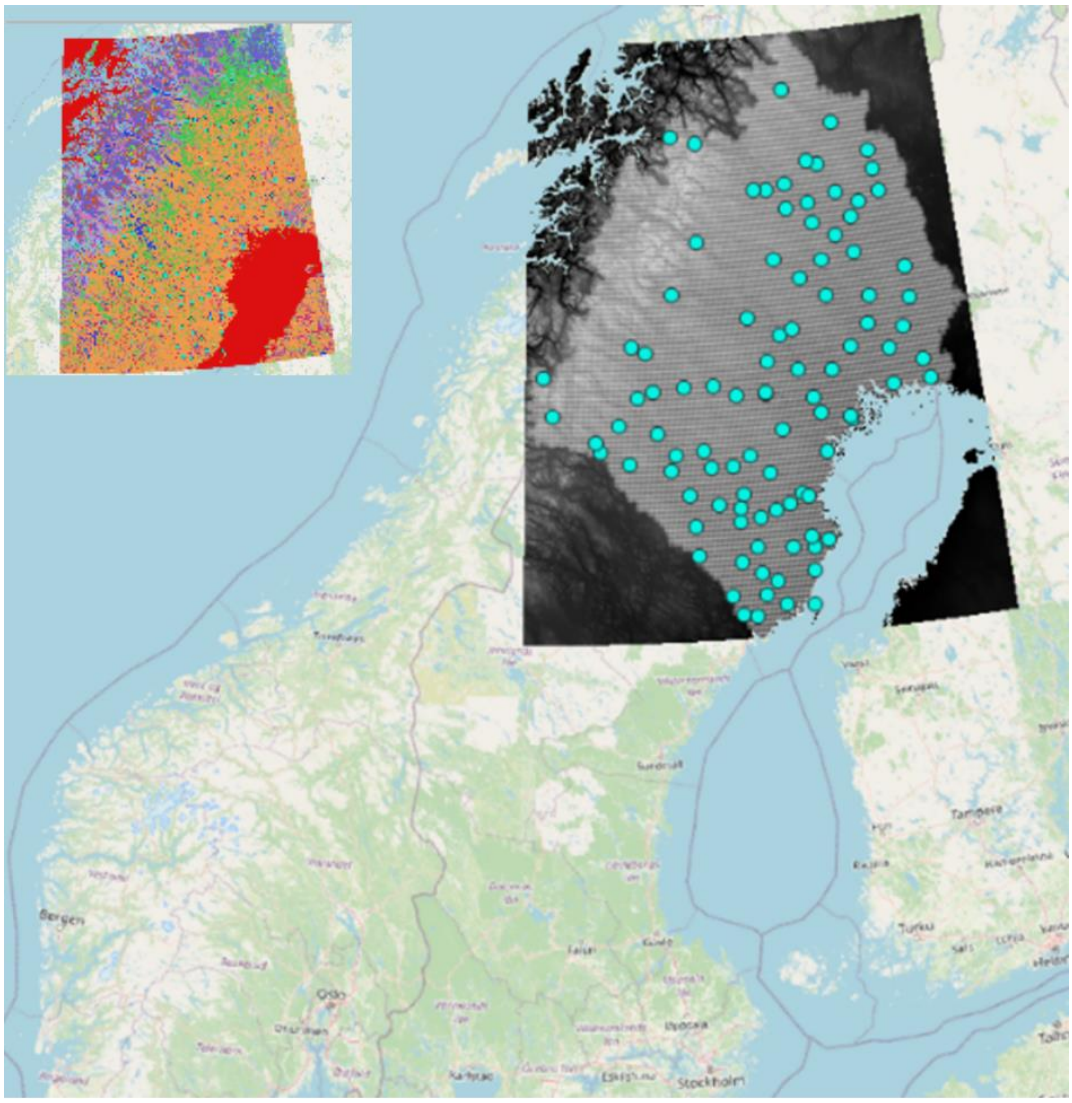

Figure 19: Northern Sweden showing the gridded HYPE model domain (in gray) at the spatial resolution of 0.05x0.05 degree and overlaying the local topography. The points in cyan are the 98 SMHI stations with daily snow depth observations. The figure extract represents the land cover provided by the ESA CCI project.

3.2.2 EnKF sensitivity to localization and error model parameters when assimilating point snow data.

A number of parameters control the performance of the Ensemble Kalman Filter (EnKF) although they are usually chosen either arbitrarily or based on expert knowledge. The main parameters in our implementation of the filter are the permitted errors and the localizations in the horizontal and vertical directions. Localizations control the distance over which an observation is permitted to have a positive covariance that diminishes to zero beyond that distance. There are also perturbation parameters in the forcing data, but those were not changed in these experiments. We used the inflow region of the Överuman reservoir in northern Sweden for which there are recorded estimates of the local runoff that flows into it. The spatial extent was extended to 6 km around the inflow region (see Figure 18). We divided the domain into two bands below and above the tree-line, which was set at 700 m above sea level.

The experiment in its entirety aims at using the modeled reservoir inflow as the target variable to investigate the impact of the three parameters named above on the assimilation of the reservoir inflow, snow water equivalent and the snow depth; as well as the impact of the

distance over which the assimilated observations impact the EnKF performance. We limit the presentation here to the parts of the investigations pertaining to the EnKF parameters. The model performance is assessed with the KGE metric, which gives further insights into the timing, bias and variability of the modeled each of the three EnKF parameters under investigation (top) when SWE above the tree-line is assimilated. The bottom are the KGE for the individual years from the investigated parameters when SWE below and above the tree-line is assimilated.

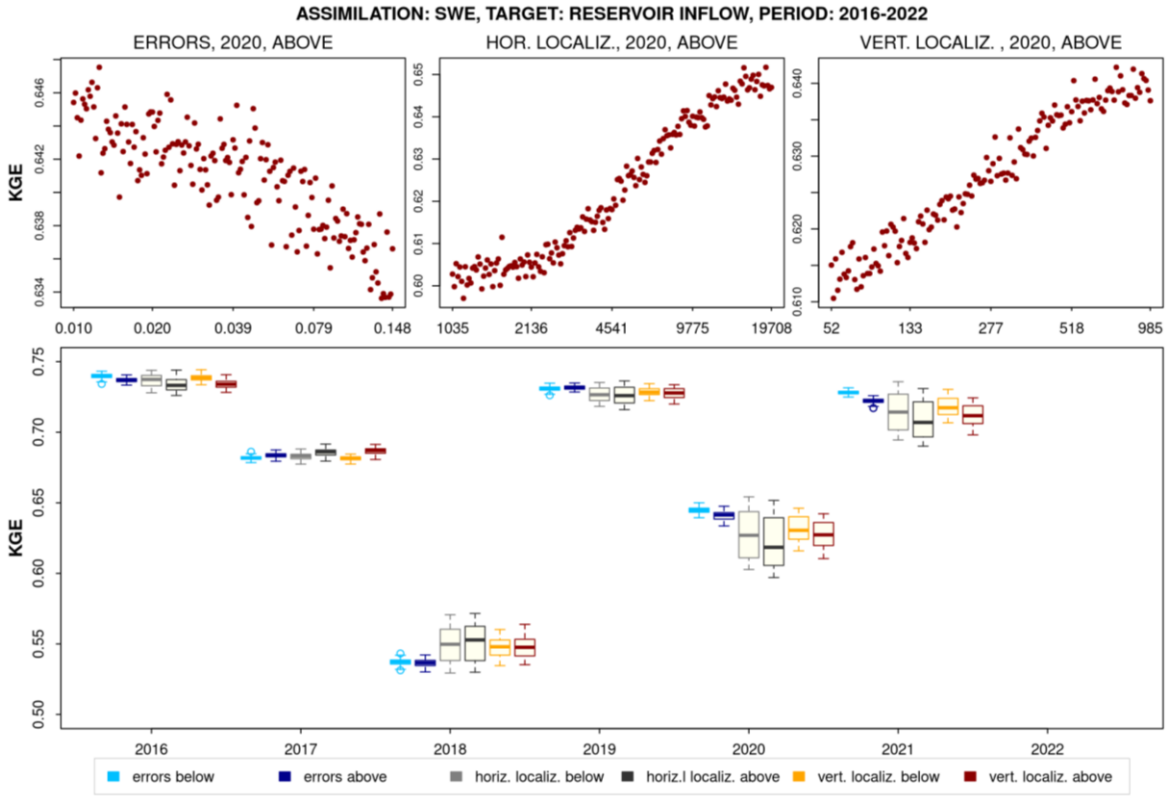

Figure 20: The KGE with each of the three EnKF parameters under investigation (top) when SWE above the tree-line is assimilated. The bottom are the KGE for the individual years from the investigated parameters when SWE below and above the tree-line is assimilated.

Figure 20 (top) shows the dependence of the KGE metric on the three EnKF-parameters during 2020 and the assimilation of the SWE above the tree line. The model performance reduces with increasing errors as would be expected. The performance increases with the localizations in both directions and tends to maxima at the upper limits. The turning points are not well defined, which reveals the need to increase the upper limits of parameters.

The bottom part of the figure shows the performance for the years in the simulation period for the assimilation of SWE below and above the tree line. There is significant inter-annual variability in the performance with the different EnKF-parameters and the spread caused by the errors (first 2 boxes) is lower than that caused by localizations (last 4 boxes). There are generally similar performances from the parameters for assimilations below and above the tree line except during 2020 and 2021 when the assimilation below the tree line resulted in higher performance than its counterpart. The results for 2022 (not shown with the displayed scale) were significantly lower than the rest and did not show the expected trends. This is possibly due to the short record length.

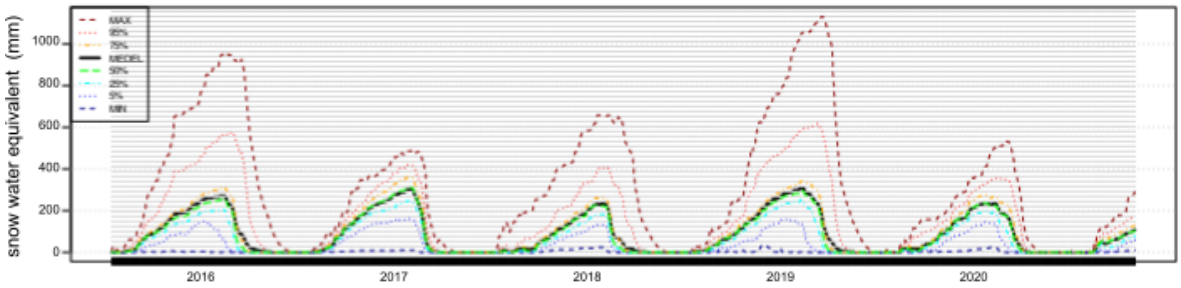

Figure 21: The aggregated snow water equivalent over the SE-northern model domain for the periods Oct 2016 to Dec 2021. The green and dark black lines indicate the median and mean, respectively.

Figure 21 shows the aggregated snow water equivalent from all the stations in the SE-northern model. Although the maximum values vary significantly over the years, the means are comparable. However, 2018 and 2020 had less than 250 mm of average snow, which was lower than the other years. The same years had low performance in the boxplots shown in Figure 20. The Överuman for which the boxplots were produced is significantly smaller than the SE-northern domain and is at a higher elevation but the latter model should give a more representative average snow amount that is free from localized effects of e.g., wind. We can therefore attribute the low model performance to the low amounts of snow during those years.

4 Conclusion

This report summarizes the development done in the CERISE project towards unified ensemble-based data assimilation framework for regional reanalysis applications. Regional reanalysis applications require an adequate treatment of soil and surface variables on a wide range temporal and spatial scales. The CERISE project aims to progress on three following topics.

- A homogenization of the analysis of snow and soil variables is one the aims of the CERISE project. At present different model variables are analysed using different methods, which is leading to the inconsistencies in the analysis and to heavy maintenance burden.
- A development of flexible ensemble-based data assimilation into the ISBA-Diff soil model and the multi-layer snow scheme is the second task of CERISE. Advanced physical models together with flexible data assimilation schemes able to handle a variety of observations from different platforms will improve quality reanalysis products of near surface variables.
- First steps are taken towards development of a consistent hydrological - meteorological forecasting system that is necessary to properly address evolution of water cycle including snow are taken in CERISE. : a) to this point, developments and assessment of snow data assimilation has been done separately in meteorological and hydrological forecasting systems but with converging methodologies to support the final implementation for the CERISE re-analysis, b) development of roadmap for implementation of a river network constrained localization to enable assimilation of streamflow observations.

An extensive comparison has been done between sEKF and EnSRKF DA scheme for inline DA assimilation of screen level 2Tm and 2RHm into ISBA-DIFF model. EnSRKF propagates increments from the screen level variables deeper into the soil. Validation has been performed over supersites against independent observations. Validation results show that analysis using the EnSRKF scheme results in a better agreement with observations than the analysis done using sEKF scheme including the temperature inversion. Enlarging the dimensionality of the control vector brings analysis closer to observations. One of the reasons is that the larger dimensionality of the control vector allows better representation of memory of the nonlinear system. In general, we have noticed that EnSRKF provides a somewhat stronger response to the daily cycle variations for deeper soil levels than the observations suggest. The reason for this behaviour is under investigation. One possible reason is an existence of systematic model errors in modelling moisture processes in HARMONIE-AROME. At the same time validation of soil variables is very challenging because soil observations are very rare. Also, validation of soil moisture analysis is more problematic than the soil temperature analysis because of heterogeneity soil types and sensitivity to predicted precipitation patterns that might differ from those that have occurred in reality.

Assimilation of remote sensing products and radiances to improve soil analysis is one of the motivations for the CERISE project. LETKF scheme has been developed to be used in the off-line Land DA. Attractive features of the LETKF is high scalability and the inherent ability to treat the footprint of the observation operator. The LETKF scheme has been extensively tested for DA in the offline multi-layer snow model and has been tried as an in-line DA scheme for screen level observations for the soil moisture analysis. Because of limited ensemble size, both vertical and horizontal localisation is introduced in order to hamper spurious correlation. LETKF is able to significantly improve snowpack in the areas of flat orography. Performance is less satisfactory in the mountainous areas where modelling of snowpack is very challenging. The spread of the LETKF ensemble, and the amplitude of the Kalman gain, is driven a lot by spread in forcing conditions. It was noticed that the spread of the LETKF scheme is substantially different in the online and offline experiments. This is partially due to different methodologies used for generation of perturbations. As it is expected the offline LETKF

scheme struggles to capture weather dependency. The diagnosis of the sanity of the Kalman gain was performed in the inline experiments and preliminary results show reasonable and intuitively sensible results. Evaluation is on-going.

The VAE framework has been tried to derive a system that is generic enough to handle a variety of situations in a unified framework. This particularly concerns the dimensionality of the control vector that currently is chosen based on affordability constraints. So far results are not satisfactory because VAE results in notably over smoothed time evolutions. Different remedies are considered to improve that behaviour. Also, the usefulness of the truly unified system is doubtful. The snowpack is heavily isolating the soil from the screen level observations. A practical solution is to derive two different systems, one is to perform analysis of soil variables when there is no snow and another to perform analysis of snow variables when snow is present.

The ensemble Kalman filter has been implemented and refined for the snow data assimilation in hydrological model HYPE. The focus was on exploring the features of the Ensemble Kalman Filtering and tuning it to the optimal performance. The sensitivity to the observation error variance and the vertical and horizontal localisation scales was evaluated. The target is the modelled river discharge that provides a measure of accumulated snow mass. The surprising discovery was that the performance of the EnKF is increasing with increasing the length scales of the localisation. The reason for this is under investigation. The work is going on with implementation of the hydrologically constrained localisation to handle upstream observations.

5 References

Blyverket J., Hamer P.D., Bertino L., Algergel C., Fairbain D, Lahoz W. A.,, 2019, “Improving soil moisture estimate over contiguous US using satellite retrievals and ensemble based data assimilation techniques”, *Remote Sensing*, **11**, 478

Charrois, L., Cosme E., Dumont M. Lafaysse M, Morin, S, Libois Q, Picard G., 2016 “On the assimilation of optical reflectance and snow depth observations in a detailed snow pack model”, *EGU The Cryosphere*, **10**, 1020-1038

Evensen, G. 1994 ”Sequential data assimilation with a non-linear quasi-geostrophic model using Mpnte Carlo methods to forecast error statistics”. Journal of Geophysical research : Oceans, 99, 10143-10162, https://doi.org/10.1029/94JC00572

Evensen, G.. Data Assimilation 2009. Springer, Berlin, Heidelberg.

El Gharamti, M., McCreight, J. L., Noh, S. J., Hoar, T. J., Rafieei Nasab, A., and Johnson, B. K. “Ensemble streamflow data assimilation using WRF-Hydro and DART: novel localization and inflation techniques applied to Hurricane Florence flooding”, Hydrol. Earth Syst. Sci., 25, 5315–5336, https://doi.org/10.5194/hess-25-5315-2021, 2021.

Lindström G., Pers C., Rosberg J., Strömqvist J., Arheimer B., 2010 “Development and testing of the HYPE (Hydrological Predictions for the Environment) water quality model for different spatial scales” . Hydrology Research.; 41 (3-4): 295–319. doi: https://doi.org/10.2166/nh.2010.007

Musuuza, J.L.; Gustafsson, D.; Pimentel, R.; Crochemore, L.; Pechlivanidis, I. 2020 “Impact of Satellite and In Situ Data Assimilation on Hydrological Predictions”. Remote Sens., 12, 811. https://doi.org/10.3390/rs12050811

Musuuza, J. L., Crochemore, L., & Pechlivanidis, I. G., 2023 “Evaluation of earth observations and in situ data assimilation for seasonal hydrological forecasting”. Water Resources Research. , 59, e2022WR033655. https://doi.org/10.1029/2022WR033655

Revel, M.; Ikeshima, D.; Yamazaki, D.; Kanae, S. A, 2019 “Physically Based Empirical Localization Method for Assimilating Synthetic SWOT Observations of a Continental-Scale River: A Case Study in the Congo Basin” . Water, 11, 829. https://doi.org/10.3390/w11040829

Document History

Version	Author(s)	Date	Changes
0.1	Jelena Bojarova, Abhishek Lodh, Jostein Blyverket, Åsmund Bakketun, Jude Musuuza, Ilaria Clemenzi, David Gustafssion, Tomas Landelius	02-12-2024	Initial version
1.0	Jelena Bojarova, Abhishek Lodh, Jostein Blyverket, Åsmund Bakketun, Jude Musuuza, Ilaria Clemenzi, David Gustafssion, Tomas Landelius	Dec and Jan 2024	Revisions due to internal reviews
1.1	Coordinator	Jan 2025	Minor formatting revisions

Internal Review History

Internal Reviewers	Date	Comments
Xiaohua Yang, DMI, Kristian H Møller, DMI and Kristina Froehlich DWD	Dec 2024	Initial version