CopERnIcus climate change Service Evolution

D6.1: Report providing a protocol for assessing the improvement in the quality in the demonstrators

Due date of deliverable	30/6/2025
Submission date	26/6/2025
File Name	CERISE-D6-1-V1.0
Work Package /Task	WP6 / Task 6.1
Organisation Responsible of Deliverable	BSC
Author name(s)	Markus Donat (BSC), Jeff Knight (MetO), Frederic Vitart (ECMWF), Constantin Ardilouze (MF), Anais Barella-Ortiz (MF), Fabrizio Baordo (DMI), Carla Cardinali (CMCC), Giovanni Conti (CMCC), Jonny Day (ECMWF), Jean-Christophe Calvet (MF), Emanuel Dutra (IPMA), Sofia Ermida (IPMA), David Fereday (MetO), Silvio Gualdi (CMCC), Stefano Materia (BSC), Gabriel Narvaez-Campo (MF), Yvan Orsolini (NILU), Lluís Palma (BSC), Daniele Peano (CMCC), Carlos Peralta Aros (DMI), Núria Pérez-Zanón (BSC), Patricia de Rosnay (ECMWF), Monalisa Sahoo (BSC), Hauke Schulz (DMI), Retish Senan (ECMWF), Tim Stockdale (ECMWF), Isabel Trigo (IPMA), Ekaterina Vorobeva (NILU), Xiaohua Yang (DMI)
Revision number	1.0
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

The CERISE project is aimed at creating innovative improvements in the Copernicus Climate Change Service (C3S). Within this framework, the focus is on the representation of the land surface in the reanalysis products and seasonal forecasts that form part of the service. This is being done through far greater exploitation of observational land surface data in atmosphere-ocean-land-cryosphere analysis than has previously been achieved. Innovative methods to assimilate these datasets have been developed in work packages (WP) 1 and 2 of CERISE and prototype reanalysis products produced (WP3, 4). In addition, the impact of these developments on the C3S seasonal prediction service, which uses reanalyses to initialise the seasonal prediction ensembles, is being assessed through the production of new sets of demonstrational seasonal hindcast simulations (WP5). WP5 will include an assessment of improvements in performance using standard metrics, but there is a need for a more complete, more in-depth, process-based evaluation of the new prototypes and demonstrators. This is the work of WP6, and this report forms part of its delivery.

Standard verification methods may not be sensitive enough to detect the often subtle changes and improvements associated with representation of land characteristics. Beyond the standard approaches applied in WP5, existing infrastructure available for making detailed assessments of the land surface in CERISE outputs is limited. In WP6, therefore, it has been necessary to develop a novel suite of tools and methodologies to fulfil this function. This is task 6.1 of WP6, and this deliverable (D6.1) reports on the outcomes of that activity. The deliverable presents a variety of methods that have been developed to evaluate the fidelity of land variables and land-atmosphere interactions in the reanalysis prototypes and seasonal prediction demonstrators. Note that this report is concerned with the development of the techniques that are to be deployed. The evaluation of CERISE developments that will be produced using these techniques will be reported later in the project.

The work produced for this deliverable has resulted in a wide range of tools and methods for in-depth assessment of prototype and demonstrator characteristics. These tools are described in detail in the report. The scope has been across a wide range of variables – as well as fundamental land-surface properties like snow cover variables, soil moisture and surface temperature, related variables such as evapotranspiration and latent heat flux, as well as atmospheric temperature, cloudiness and circulation, have been considered. For many of these variables, the development and cataloguing of observational datasets in WP7 has been essential to producing an effective toolkit for evaluation.

There is a similar breadth in methodological approaches. Methods to assess skill and reliability of snow cover forecasts are complemented by tools to examine snow cover distributions and snow-atmosphere coupling considering the direction of causality. Off-line hydrological simulation using CERISE inputs has also been developed to provide a novel way of assessing the performance of integrated land-surface properties in simulations. Assessment methods for soil moisture have highlighted the role of observational uncertainty. Novel ways of analysing the performance of newly-produced systems in terms of soil moisture-atmosphere coupling, at scales from local and global, have also been produced. Some of these techniques have harnessed machine learning to try to separate the roles of the land surface and the atmosphere in generating local climate extremes. A range of further methodologies and tools have also been developed and are detailed in the report. For example, there are tools to examine what seasonal-scale error growth tells us about seasonal prediction performance and assessments of consistency and trends in reanalyses and seasonal hindcasts.

The collection of methods developed here create an experimental toolkit for the assessment of the innovations in reanalyses and seasonal predictions that are being created in CERISE. As mentioned above, these tools will be deployed for this function later in the project. Since

the project has had to conceive and develop these tools as research and development activity, there is no guarantee that they will be universally successful in providing clear evidence of improvement (or otherwise) in CERISE outputs. Nevertheless, by developing a broad range of techniques looking at different products and variables, gives the best chance of being able to clearly determine the benefits of the main innovations in the CERISE project.

Table of Contents

2 Introduction

Interactions between the land and the atmosphere are important modulators of climate variations, and therefore sources of predictability. A key goal of the CERISE project is to develop new methods for the assimilation of land surface observations to enhance the representation of time-varying land surface properties in climate reanalysis (https://www.cerise-project.eu/). CERISE aims to create prototype reanalysis products and test the value of these products by using them to initialise seasonal predictions as a demonstration of potential benefit to the C3S service. Key to the implementation of CERISE is the evaluation of potential improvements in these newly-developed reanalysis prototypes and seasonal demonstrators. This is the function of work package (WP) 6 ('Evaluation and exploitation of demonstrator results for future C3S implementations'). The initial phase of activity in WP6 (Task 6.1 – 'Develop techniques and methodologies for evaluating the increased fidelity of land surface processes in the prototypes and demonstrators') concerns the creation of the necessary tools to attempt to evaluate improvements in quality. These tools will be applied to the new reanalysis prototypes and seasonal prediction demonstrators produced by CERISE in Task 6.3, work which is scheduled to occur later in the project. This deliverable (D6.1) documents the collection of evaluation methodologies and tools to assess potential future improvements. This is in addition to the more traditional methods of seasonal forecast evaluation (e.g. probabilistic skill scores) which will be performed in WP5. Because these can lack sensitivity in robustly detecting changes in the fidelity of predictions, it is necessary to develop a range of other techniques to understand the impact of system changes brought about by improved land surface representation. The set of tools and methodologies described here establish a framework and a protocol for assessing the new reanalysis and seasonal forecasts in CERISE.

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, research and innovation (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 deliverable

This deliverable presents a variety of techniques and methodologies for evaluating the increased fidelity of land surface processes in the newly developed reanalysis and seasonal prediction prototypes and demonstrators developed within CERISE. The objective is to develop an innovative set of tools and diagnostics that allow the evaluation of different characteristics and processes in relation to land state and land-atmosphere interactions. These tools are being developed and tested on current reanalyses and C3S seasonal prediction systems (so-called 'phase zero' demonstrators), in preparation for being applied to the new demonstrators and prototypes developed within CERISE.

2.2.2 Work performed in this deliverable

This deliverable is the output of WP6 Task 6.1, which focuses on developing new techniques and methodologies to evaluate the improved accuracy of land surface processes within new reanalysis prototypes and seasonal prediction demonstrators. The central goal is to establish a robust framework for assessing these improvements and identifying suitable datasets for verification. This task develops a variety of methods for evaluation, moving beyond standard verification methods (as applied in WP5) that may not be sensitive enough to detect the sometimes subtle effects of the land-atmosphere interactions. These methods include the evaluation of land states and land-atmosphere interactions associated with soil moisture and snow cover, the verification of river discharge, understanding error growth associated with land-atmosphere coupling in the initialised predictions, evaluating the consistency of hindcasts and forecasts, and trends in the hindcasts. The task also focuses on determining suitable datasets for the evaluation of the land-related characteristics, and on identifying so-called 'windows of opportunity' when the land state provides enhanced predictability.

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

2.3 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 Tools for testing improvements in the reanalysis prototypes and seasonal prediction demonstrators

3.1 Assessing the fidelity and reliability of snow re-forecasts and reanalyses

3.1.1 Relevant variables and evaluation of snow re-forecasts

Snow variables. Most of the satellite snow observations provided by WP7 are available as snow cover fraction (SCF), while the majority of forecast systems provide Snow Water Equivalent (SWE) or snow depth (SD) as the prognostic variables. SCF is calculated following a conversion rule that is model dependent. For example, in the ECMWF Phase 0 demonstrator, a linear conversion rule is used, with a SD threshold of 0.1 m implying 100% snow cover, while in Phase 1 a more complex conversion rule will be employed based on Niu and Yang (2007). The latter is already in use in the CMCC Phase 0. Once converted, the SCF in planned demonstrators will be compared directly to the satellite SCF observations in the WP7 repository, namely the IMS (2004 - 2022), and the two ESA-CCI products AVHRR (2000 - 2018) and CryoClim (1982 - 2019), or to re-analyses. Comparison to these reference datasets will be the basis for assessing improvement in SCF.

Forecast fidelity and skill assessment. In a first instance, the bias and root-mean square error can be calculated to assess the fidelity. The impact of land data assimilation (DA) (through initial conditions) on the forecast skill in different project phases (or in various experiments in the same phase) can be assessed via the r-square metric (e.g., Li et al., 2019). In this case, r is the anomaly correlation coefficient between the ensemble-mean re-forecasts and observations (reanalyses). The difference in r^2 (with the sign of r) between different phases of the demonstrators (or between different experiments in the same phase) will indicate regions where differences in land DA are resulting in either improved or reduced forecast skill. Such diagnostics will be applied in short (10-days) to long (seasonal mean) time windows based on the purpose of the analysis. In addition, the Spatial Probability Score (SPS; Goessling and Jung, 2018) has been added to ECMWF's Coupled Ensemble Predictions Diagnostics (CEPDIAG) package and will be used to evaluate snow cover in CERISE seasonal forecast demonstrators. It is a spatial analogue of the Continuous Ranked Probability Score (CRPS) and is defined as $SPS = \int \{Pf(x) - Po(x)\}^2 \, dV$, where Pf is the probability of snow cover and Po is 1 or 0 depending on whether snow is present or not. As far as we know, this is the first time that the SPS has been applied to snow-cover, having previously been used to evaluate sea ice cover.

Reliability diagrams for snow forecasts. For a probabilistic forecasting system, it is important to assess whether its predictions are reliable, i.e. that the forecasted probability of a binary event matches its actual frequency of occurrence. Reliability diagrams visualize such information for the entire range of forecast probabilities. A forecasting system is perfectly reliable when the data points in a reliability diagram lie on a diagonal. We can assign reliability categories based on the slope of the best-fit reliability line and the uncertainty associated with it following Weisheimer and Palmer (2014): 5 – perfect (green), 4 – useful (blue), 3 – marginally useful (yellow), 2 – not useful (orange), 1 – dangerously useless (red). A tool for reliability assessment of snow forecasts has been developed and evaluated on phase zero seasonal hindcasts for different regions in the Northern Hemisphere (Figure 1). It can also be applied to shorter time windows. The uncertainty range is defined as the 90% confidence interval of best-fit slopes obtained via 1000 bootstrap tests with replacement applied to the order of years in re-forecasts and observations. By comparing such reliability maps for each CERISE model

and demonstrator phases, regions with improved/reduced reliability can be identified. The reliability assessment can be carried out against snow variables in reanalyses as well as against snow observations from the satellites provided by WP7.

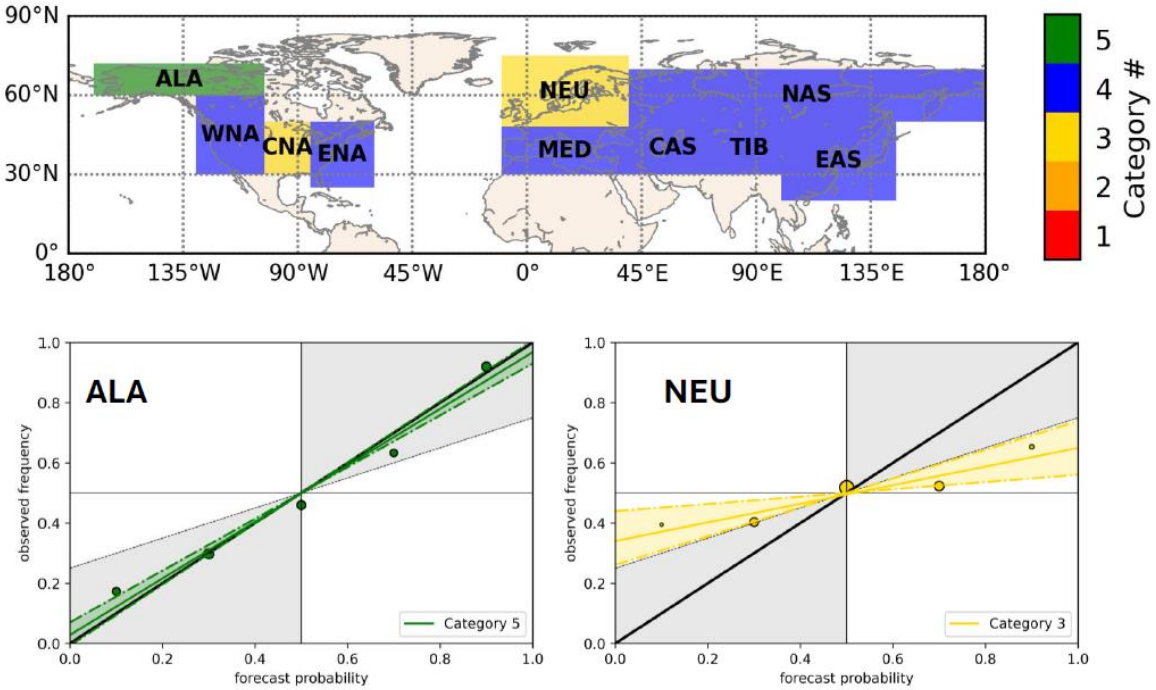

Figure 1: Map (top) of reliability categories for the ECMWF phase 0 re-forecasts against ERA5, for SWE in DJF 1993-2022. Here, the binary event E corresponds to a snow anomaly above the median. Also shown (bottom) are two reliability diagrams for two specific regions (Alaska and Northern Europe). Color codes for different categories are explained in the text.

3.1.2 Snow processes and snow-atmosphere coupling

Snow phenology. To compare snow phenology in demonstrators and re-analyses, methods to derive a series of indicators have been developed. Such indicators include the duration of snow melt or accumulation, or the day of snow onset or snow ending.

Snow-atmosphere/coupling. A classical approach to analyze the land-atmosphere coupling is to look at the r-square metric as a function of lead time, where r is the anomaly correlation coefficient between the near-surface air temperature and snow depth or SWE (Figure 2). Regions with the strong negative correlation are also called the “cold spots” of snow-atmosphere coupling and are found near the snow transition line (Xu and Dirmeyer, 2011; Li et al., 2019). Methods to identify regions of snow-atmosphere coupling have been developed and will be applied to the models in the new demonstrator phases to identify where it is realistic and where it is under-represented or exaggerated.

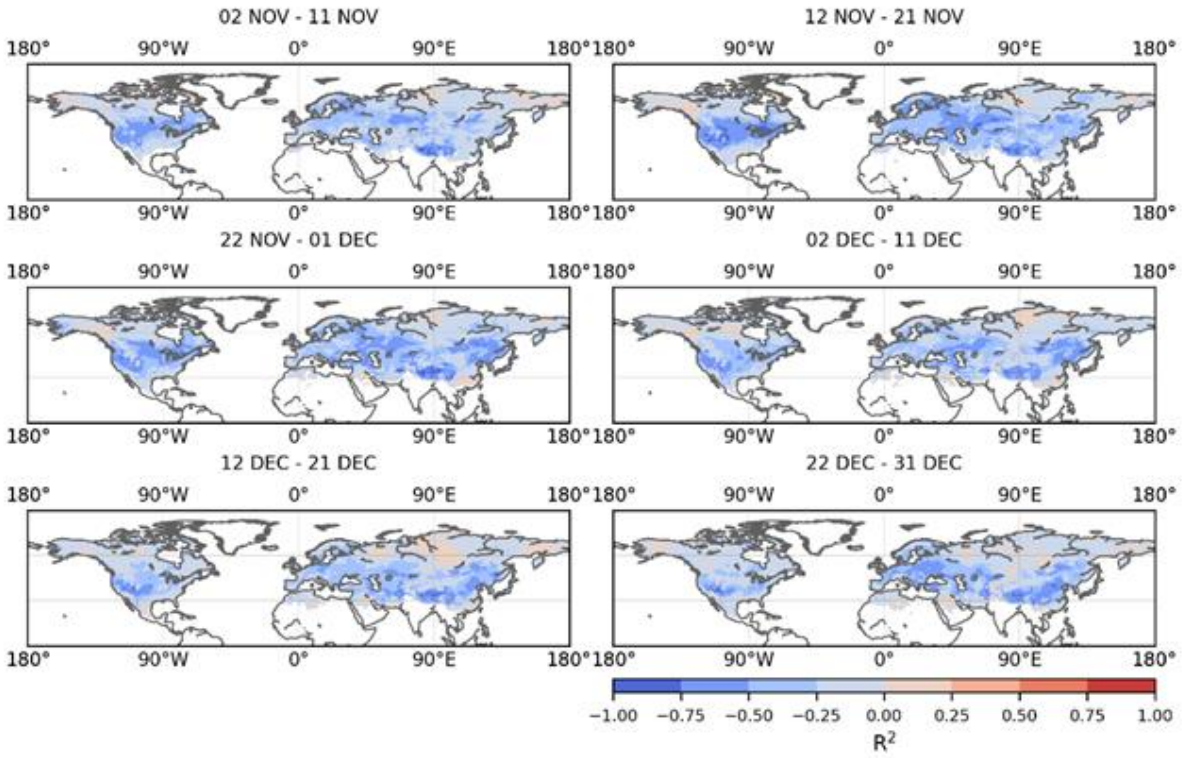

Figure 2: The correlation-square (r2) with the sign of r between the SWE (m WE) and temperature at 2m (K) as a function of lead time. Shown for the ECMWF phase 0 re-forecasts in 1993 – 2022 initialized on 1 November in six 10-day windows (lead 0: 2–11 November, lead 1: 12–21 November etc).

Coupling strength. Similar to what is done for soil moisture, a coupling strength has been inferred from the characteristics of the forecast ensemble. In its original form, this so-called Ω

$$\Omega = \frac{N\bar{\sigma}^2 - \sigma^2}{(N-1)\sigma^2}$$

diagnostics necessitates comparing twin experiments (Koster et al., 2006; Xu and Dirmeyer, 2011) but is close to potential predictability. Ω is associated with the mean value of the anomaly cross correlation coefficient and the average variance. For the snow forecast assessment, we will analyze forecasts during snow-accumulation and melting periods. By comparing Ω diagnostics in different demonstrator phases, regions with improved or reduced consistency between the ensemble members can be identified or compared between models.

Causality inference. While correlation analysis can find a statistical linear association between variables, it cannot infer the direction of causation. The Liang–Kleeman information flow approach can evaluate the cause in dynamical systems and is calculated as

$$T_{2\to1} = \frac{C_{11}C_{12}C_{2,d1} - C_{12}^2 C_{1,d1}}{C_{11}^2 C_{22} - C_{11}C_{12}^2},$$

where C_{ij} is the covariance between variables X_i and X_j, and $C_{i,dj}$ is the covariance between X_i and the time derivative of X_j using the Euler forward scheme.

This diagnostic tool has been applied before to the snow-temperature coupling problem in S2S prediction (Komatsu et al. 2023, Takaya et al. 2024), but not yet in seasonal forecasting. An example is shown on Figure 3 to infer how surface temperature and snow depth influence each other. Moreover, the methodology to infer causal linkage will be applied between the snow and soil moisture i.e., to look at hydrological effects in spring, or between snow spatial gradients and planetary waves, i.e., to link their forcing to land-sea temperature contrasts.

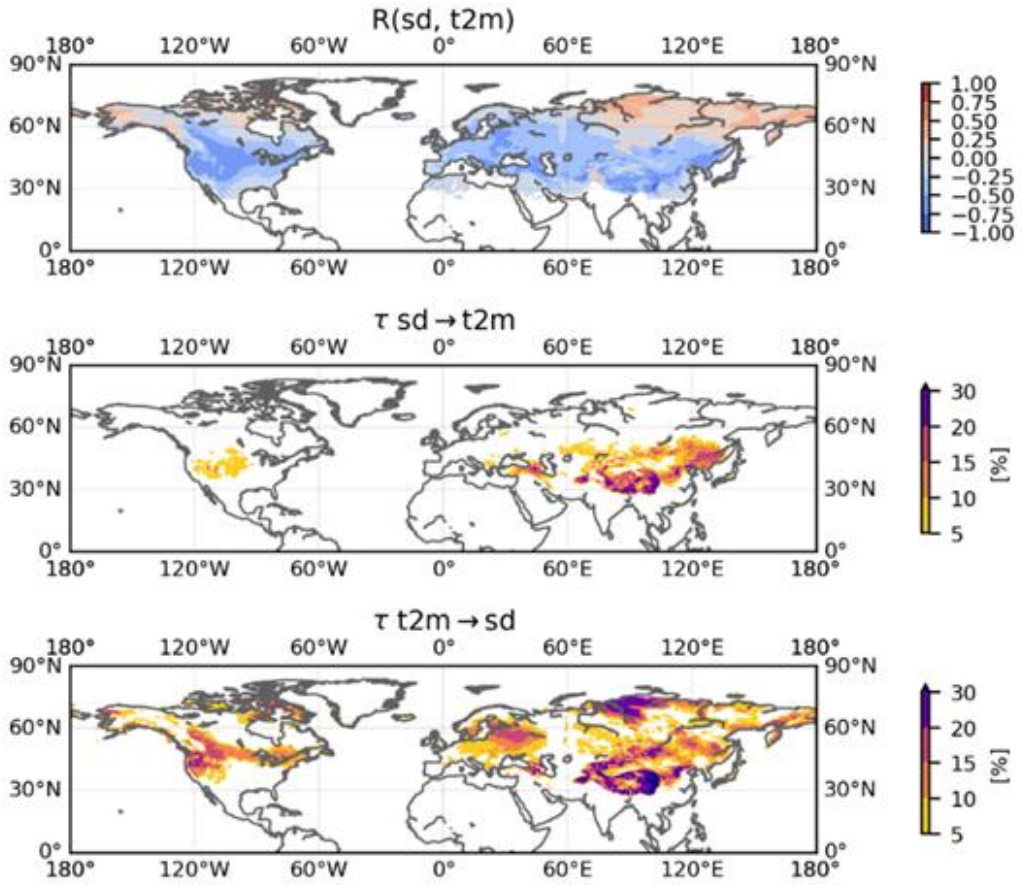

Figure 3: Comparison between the correlation and information flow shown on example of the SWE (m WE) and temperature at 2m (K) variables in the ECMWF phase 0 demonstrator. a) correlation coefficient, b) information flow from the snow depth to the temperature, c) information flow from the temperature to the snow depth. The comparison is shown for the re-forecasts initialized on 1 November initial date assessed over the first month (0-lead) in 2000 – 2019.

Using these diagnostics, we can infer causality relationships in different models and demonstrator phases and identify the impact of new, improved data assimilation approaches.

3.1.3 Verification of snow cover from reanalyses

Spatial verification techniques aim to quantify differences in field spatial structure for weather variables over spatial domains, to provide information on error that also account for time-space uncertainties. Measuring errors in snow coverage is particularly challenging, since winter precipitation measurements can show large differences between different observing networks where the exact values depend upon regional snowfall characteristics. In this case, we compare high resolution model output with satellite estimates which are also highly uncertain. For this reason, we use neighbourhood-based or fuzzy verification techniques that aim to relax requirements for exact positioning and account for time-space uncertainty. It is widely used for evaluation of forecast skill under different thresholds in spatial windows of increasing size.

The fractional skill score (FSS, Roberts and Lean, 2008) is a neighbourhood verification metric based on the probability of the occurrence of an event in different spatial windows. In this case, we use the binary variable bin_snow, and there is only one threshold available. In Figure 4 we show a case study for the period Nov 2015 to May 2016.

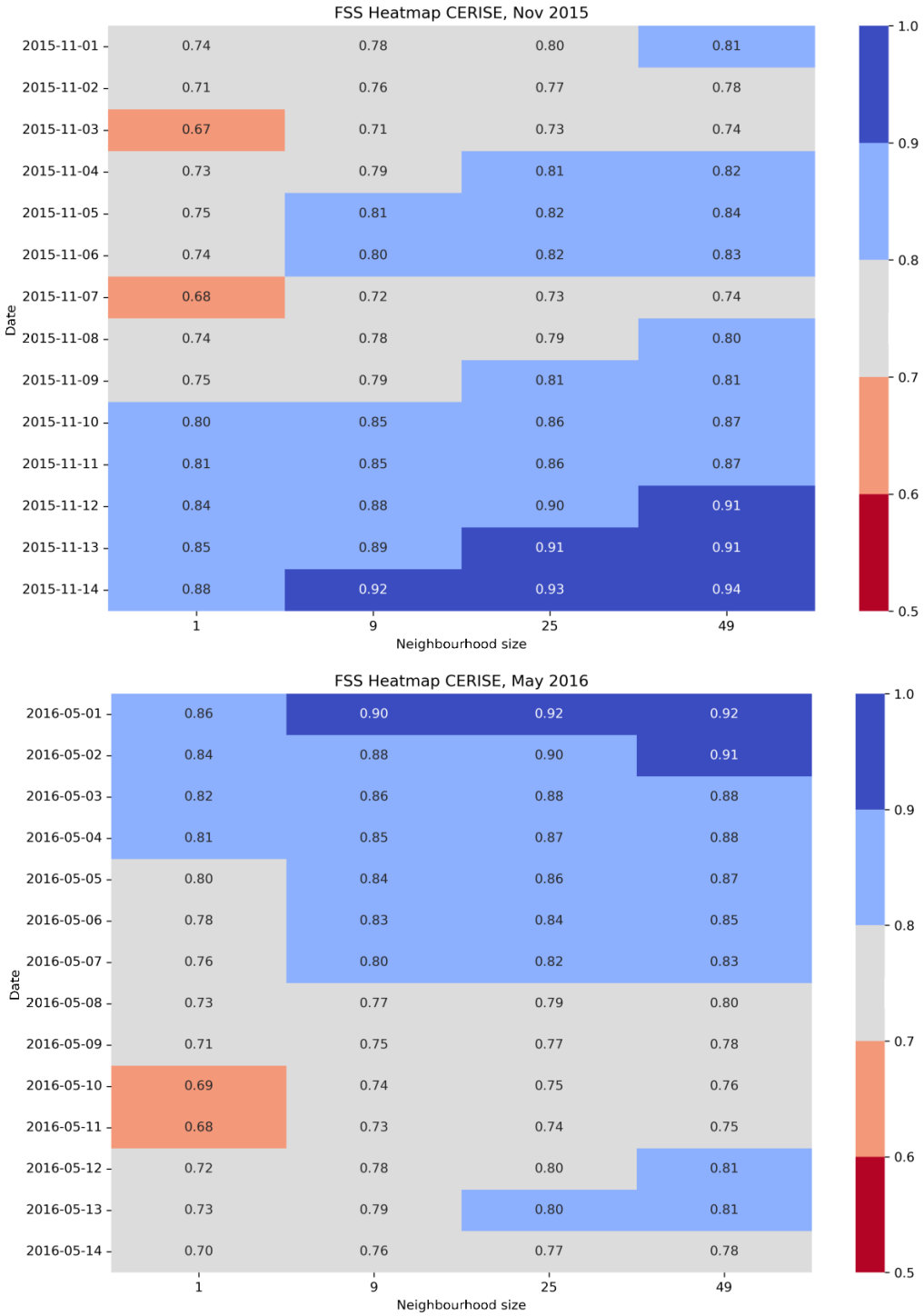

Figure 4: Heatmap of Fraction Skill Score (FSS) of binary snow for the CERISE demonstrator (based on the CARRA-Land-Pv1 data set) against IMS data. The top row corresponds to the first 15 days of November 2015 and the bottom row to the first 15 days of May 2016. All domains were mapped to the CARRA1 East grid before calculating FSS, unit of neighbourhood size is the grid size, using the resolution of CARRA1 (2.5 km).

The methodology has been tested on data from the CARRA-Land-Pv1 system. The input is in the form of analysis/observation data in Zarr format that is converted to netcdf format and then processed by the grid_stat tool of the MET package. Other reanalyses, like CARRA1 and ERA5, have also been compared against IMS data. By comparing FSS scores for different models and demonstrator phases the reliability of the spatial distribution of snow on different times of the year can be identified.

3.2 Hydrological evaluation

3.2.1 Hydrological study unit

In hydrology, the study unit where the hydrological processes yield streamflow is the drainage basin, also known as the catchment or watershed. The National Oceanic and Atmospheric Administration (NOAA) defines a drainage basin as “a land area that channels rainfall and snowmelt to creeks, streams, and rivers, and eventually to outflow points such as reservoirs, bays, and the ocean”. Taking the drainage basin as the hydrological unit, we developed a hydro-evaluation system at the basin level. This system allows any hydrological variable, such as soil moisture, rainfall, snow fraction, and runoff, to be aggregated and evaluated against observations or other models' outputs.

3.2.2 Hydro-evaluation system

A hydrological analysis system has been produced that allows the assessment of the impact from new reanalysis prototypes on seasonal forecasting. The hydrological analysis can be driven by meteorological outputs from any CERISE partner model (Figure 5). Streamflow is a widely observed quantity and is potentially a sensitive indicator of land surface conditions in response to basin rainfall and/or snowmelt. We have applied the system to evaluate streamflow derived from phase zero demonstrators so far (e.g. Narváez-Campo and Ardilouze, 2024). Predicted streamflow is assessed against the observation datasets presented in Table 1. Meanwhile, other hydrological variables can also be evaluated using the same system.

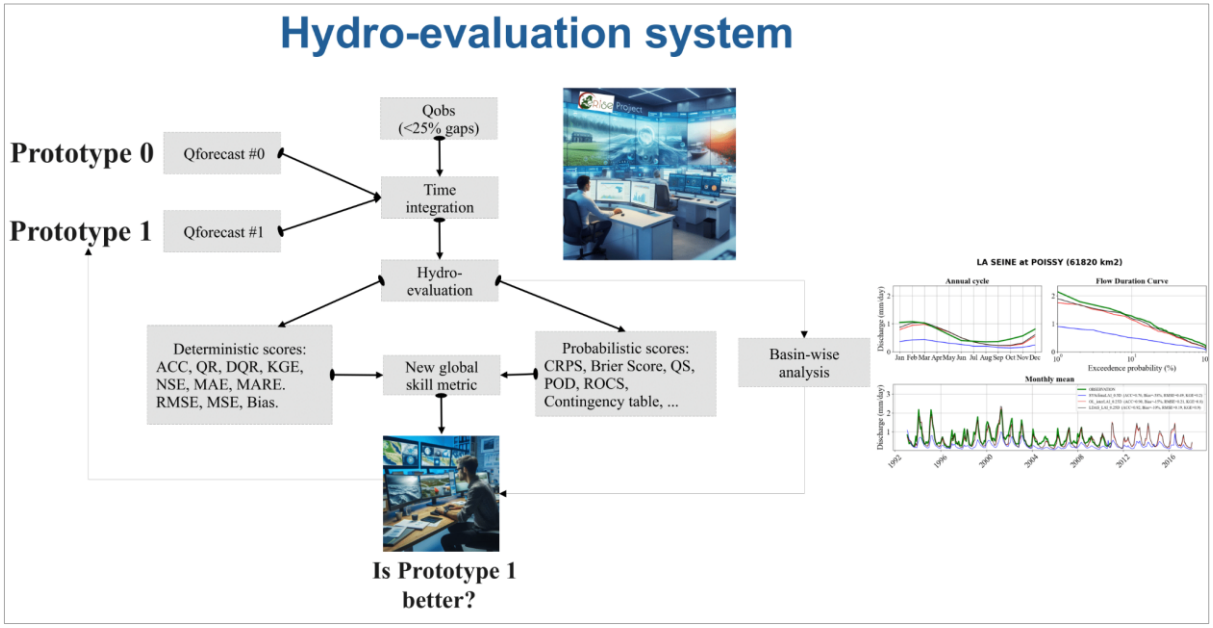

Figure 5: Hydro-evaluation system flow-chart schematic.

Dataset	Region	Reference
GRDC: Global Runoff Data Centre	Global	http://www.bafg.de/GRDC/EN/Home/homepage_node.html
USGS: United States Geological Survey	United States	http://waterdata.usgs.gov/nwis/sw
HYDAT: National Water Data Archive	Canada	https://collaboration.cmc.ec.gc.ca/cmc/hydrometrics/www/
French Hydro database	France	http://www.eaufrance.fr
Spanish Hydro database	Spain	http://ceh-flumen64.cedex.es/anuarioaforos/default.asp
HidroWeb	Brazil	http://www.snirh.gov.br/hidroweb/
R-ArcticNet	Northern High Latitudes	http://www.r-arcticnet.sr.unh.edu/v4.0/AllData/index.html
Australian Bureau of Meteorology	Australia	http://www.bom.gov.au/metadata/19115/ANZCW0503900339
China Hydrology Data Project	China	Henck et al., 2011
HyBAm	Amazon basin	https://hybam.obs-mip.fr/

Table 1: Streamflow observed datasets.

The hydro-evaluation system can be divided into the following main components:

Data Reading. Daily streamflow ensemble hindcasts are read from multiple files and stored in one multidimensional array $N_{time} \times N_{members} \times N_{computational-cells}$.

Data filtering. The full observed database is filtered to select flow-gauge stations with less than 25% of missing data (per season, month, or full data series) and a basin area lower than 6000 km^2. The observations are stored in an array of size $N_{time} \times N_{stations}$.

Localisation of flow-gauge stations. To compare observed and simulated discharges, one must first localise the gauge station within the river network of the model. This procedure is done through the methodology proposed in Munier and Decharme (2022), obtaining a new streamflow forecasted array of size $N_{time} \times N_{stations} \times N_{members}$ consistent with the observed array.

Time integration. Depending on the needs, the evaluation can be done in the native time resolution (daily or hourly) or performed monthly or quarterly. At this step, the ensemble mean is also computed to generate a new array of size $N_{time-aggreg} \times N_{stations}$ employed for deterministic metrics computation.

Scores against observations. Metrics computation is done using the evalhyd tool, whose documentation is available at https://hydrogr.github.io/evalhyd/python/. The tool allows computing 12 deterministic and 23 probabilistic scores.

Skill metrics between prototypes. The system evaluates where the new prototype was better or worse than the reference prototype. We apply the generic local skill metrics in Table 2 for a basin-by-basin comparison. Note that these generic skill metrics can be based on any scores computed in the previous bullet.

Name	Equation	Description
Absolute Skill Score	$\lvert \text{Score}_0 - \text{Score}_{\text{perfect}} \rvert - \lvert \text{Score}_1 - \text{Score}_{\text{perfect}} \rvert$	ABS ranges $(-\infty, 1]$ and RES ranges $(-\infty, \infty)$. It compares the new prototype against the previous or reference prototype. perfect skill: RES $= 1$ (ABS $= \lvert \text{Sc}_{\text{off}} - \text{Sc}_{\text{perf}} \rvert$). no skill: RES $= 0$ (ABS $= 0$). skill degradation: RES < 0 (ABS < 0). **Note:** Any deterministic or probabilistic score can be used. ABS/RES is the magnitude/fraction of the score improvement (or degradation for negative values).
Relative skill score	$1 - \frac{\text{Score}_1 - \text{Score}_{\text{perfect}}}{\text{Score}_0 - \text{Score}_{\text{perfect}}}$	

Table 2: Generic local skill scores used to compare seasonal streamflow forecast from prototype 0 against prototype 1 basin by basin (table adapted from Narváez-Campo and Ardilouze, 2024).

Besides the scores in Table 2, we propose a new global skill metric to evaluate the additional skill of the new demonstrator (relative to the reference) for a certain region instead of basin by basin. For a given region or set of basins, the new metric (Global Weighted Skill) compares the cumulative distribution function (CDF) of a certain score between the reference and new prototypes. When the CDF_{new} is more concentrated at high performance score values than the $CDF_{reference}$, it means that more basins with flow-gauge stations show higher scores. Then, the Global Weighted Skill GWS allows to objectively compare the CDF_{new} and $CDF_{reference}$ to have an accurate view of the aditional skill of the new prototype for a certain region or set of hydrologicaly similar basins. For any positively or negatively oriented score S (e.g, ACC, KGE, NSE, etc.), for which S_{low} is considered a low performance value, the GWS reads:

$$GWS = \frac{n+1}{S_{optimal} - S_{low}} \int_{S_{low}}^{S_{optimal}} (CDF_{new} - CDF_{reference}) \times (S - S_{low})^n \, dS$$

where

$GWS = 100\%$: perfect skill

$GWS \approx 0\%$: No skill

$GWS < 0\%$: less accurate than reference prototype (or climatology/persistence)

The parameter n, can be arbitrary set to $0, 1, 2$ or 3 to give different weights to different scores magnitudes. The greater this parameter, the higher the weight given to the scores near the optimal one. In other words, the higher n, the more exigent we are with the performance of the new prototype in relation to the reference prototype.

3.3 Evaluation of soil moisture variability and effects on the atmosphere

3.3.1 Relevant variables and limitations

To facilitate the evaluation of soil moisture and land-atmosphere feedbacks, a number of datasets have been considered as described in this section. In particular, the relation between soil moisture, precipitation and turbulent energy fluxes at the surface is assessed using the following datasets that have also been added to existing ECMWF's Coupled Ensemble Predictions Diagnostics (CEPDIAG) evaluation toolkits:

- Surface and root-zone soil moisture (daily and monthly mean) from GLEAM (v3.7; Miralles et al., 2011) and ERA5-land (Muñoz-Sabater, 2021).
- Berkeley Earth Surface Temperatures (BEST; Rohde et al. 2013), which provides spatially complete monthly station-based air-temperature estimates over land.
- Version 2.3 of the Global Precipitation Climatology Project (GPCP; Adler et al. 2018).

These and other datasets used for the evaluation of CERISE reanalysis prototypes and seasonal forecast demonstrators are summarised in Table 3.

Satellite Land Surface Temperature (LST) and in particular its daily amplitude is closely related to the partitioning of available energy at the surface. We use high frequency (hourly) LST data estimated from Meteosat Second Generation (MSG) observations by the Land Surface Analysis Satellite Application Facility (LSA SAF) LST (Trigo et al., 2021). LSA SAF LST is available under clear sky conditions only. While it may be affected by cloud contamination, or by uncertainties in atmospheric or emissivity correction (see Trigo et al., 2021), data aggregation in space and time greatly smooths those errors: LST (originally on a geostationary projection; 3 km at the sub-satellite point) is regridded to 0.25º; the hourly LST estimates are used to compute the monthly mean diurnal cycle.

It is important to note that for several variables relevant for the evaluation of soil moisture and its effects on the atmosphere, the availability of *in situ* observations is severely limited. This affects soil moisture data itself, but also observations of heat fluxes and evapotranspiration. Some of the widely used observation-based datasets of these variables are therefore based on relatively simple land models driven by atmospheric boundary conditions provided by atmospheric reanalyses (being a numerical model output itself). We therefore performed a comprehensive analysis of the robustness of the relevant variables across a variety of observations-based datasets, and how potential data uncertainties affect not only our knowledge of the spatio-temporal variations of soil moisture (see below), but also of the coupling relationships between land and atmosphere (see following sub-section 3.3.2).

Table 3 lists the observational datasets used in this analysis. Detailed inspection of the data indicates that the JRA55 surface soil moisture data shows abnormally low values on the first day of each month, indicating a potential artefact in the dataset (not shown). Similarly, both versions of the GLEAM dataset (3.7b and 3.8a) exhibit anomalously low values for evapotranspiration and potential evapotranspiration on June 6 and 7 each year (not shown here). Regional discrepancies are also evident, with poor correlations among the observational datasets found in the case of evapotranspiration from FLUXCOM in Northern Europe (NEU), SiTHv2 in Western Central Europe (WCE), and in the case of potential evapotranspiration from GLEAM 3.7b in the Mediterranean (MED, e.g. Figure 6).

Data	**Data type**	**Resolution**	**Variables**
GPCP	Satellite & In Situ	2.5º	Monthly gridded precipitation
CLARA	Satellite	0.25º	Cloud fraction (CM SAF data record)
BEST	Gridded In-situ	1°	2m air-temperature over land
LSA SAF LST	Satellite	0.05º	Land Surface Temperature (clear sky)
ESA-CCI	Satellite	0.25 °	Surface soil moisture

Data	Data type	Resolution	Variables
FLUXCOM	Gridded in situ	0.5 °	Latent heat flux, **Evapotranspiration**
E-OBS	Gridded in situ	0.25 °	Surface air temperature, Maximum air temperature, Minimum air temperature, **Potential evapotranspiration**
GLEAM	Model	0.25 °	Soil moisture, Evapotranspiration, Potential evapotranspiration
SiTHv2	Model	0.1°	Soil moisture, Evapotranspiration
ERA5/ ERA5-land	Reanalysis	0.25° /0.1°	Soil moisture, Latent heat flux, Net surface longwave radiation, Net surface shortwave radiation, **Evapotranspiration**, Surface air temperature, Maximum air temperature, Minimum air temperature, **Potential evapotranspiration**
MERRA2	Reanalysis	0.5 °	Soil moisture, Evapotranspiration, Surface air temperature, Maximum air temperature, Minimum air temperature, **Potential evapotranspiration**
JRA55	Reanalysis	0.5 °	Soil moisture, Evapotranspiration, Surface air temperature, Maximum air temperature, Minimum air temperature, **Potential evapotranspiration**

Table 3: Observational datasets used in the prototype evaluation. Variable names in **bold** indicate that they are not provided as such by the dataset but derived from other provided variables**.** For analysis involving several datasets (e.g. calculating regressions or correlations between variables) these are remapped on a common grid, and similar for model evaluations these should be remapped on a common grid, or averaged over similar larger regions.

While exploring the uncertainties across different observational-based datasets, it is found that the summer (JJA) soil moisture data are in good agreement with each other across different datasets (values of correlation coefficient among all the datasets are higher than 0.7 for all the European regions) even though there are substantial differences in the magnitudes across the datasets (see Figure 6 a, b). While analysing the surface soil moisture and root-zone soil moisture data, we found that the pattern of relation of these soil moisture data with other variables (i.e., evapotranspiration, surface temperature, and potential evapotranspiration) were similar, while the relationships with atmospheric variables are much stronger in case of surface soil moisture than root-zone soil moisture. Therefore, we are not showing the results from root-zone soil moisture here. However, the differences among datasets of either actual or potential evapotranspiration across Western and Central Europe are quite considerable (Figure 6).

This analysis illustrates that there are non-negligible differences in the actual values of the different variables. When considering the time series for specific regions, surface soil moisture shows mostly significant correlations between the different datasets (Figure 6b). In contrast, evapotranspiration shows low or even negative correlations between some datasets, pointing to important inconsistencies in their temporal evolution. For example, the evapotranspiration SiTHv2 is weakly correlated with all the datasets except GLEAM 3.8a and JRA55. In addition, JRA55 indicates negative correlations with ERA5, FLUXCOM, and MERRA2 and no

correlation with ERA5-Land and weak correlation with GLEAM (3.7b, 3.8a) data. Similarly, potential evapotranspiration from JRA55 shows weak correlation with all other datasets, except GLEAM (3.7b, 3.8a) data. All these inconsistencies in the correlation matrices are seen in the spatial maps of land-atmospheric coupling as well (Figure 7).

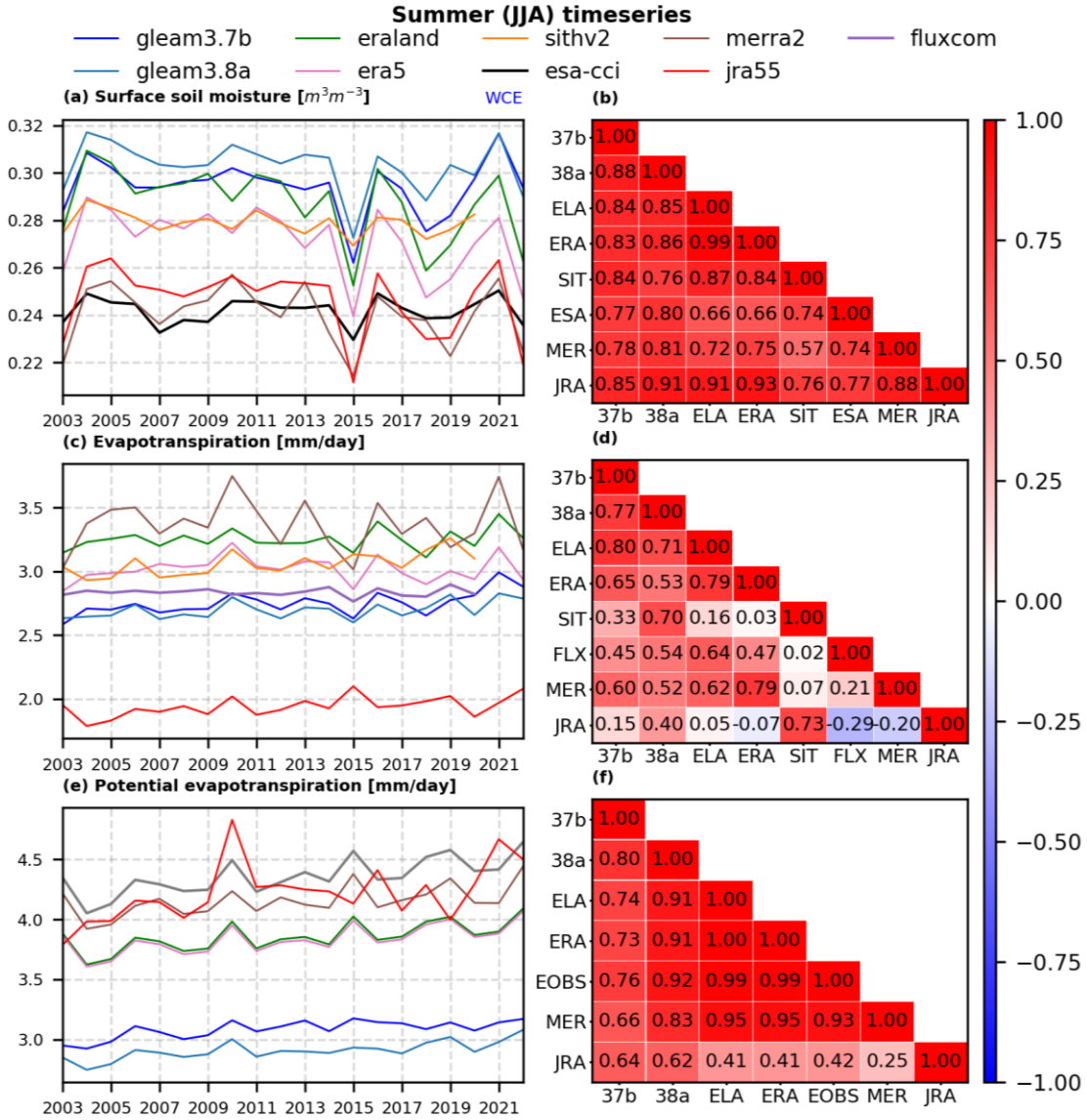

Figure 6: Summer (JJA) time series (on the left) for (a) surface soil moisture, (c) evapotranspiration, and (e) potential evapotranspiration and corresponding correlation matrices (on the right; (b), (d), (f)) for WCE. The E-OBS, FLUXCOM, and ESA-CCI datasets were regridded to the common FLUXCOM grid at 0.5° resolution, while the SiTHv2, ERA5, and ERA5-Land datasets were regridded to the ERA5 grid at 0.25° resolution.

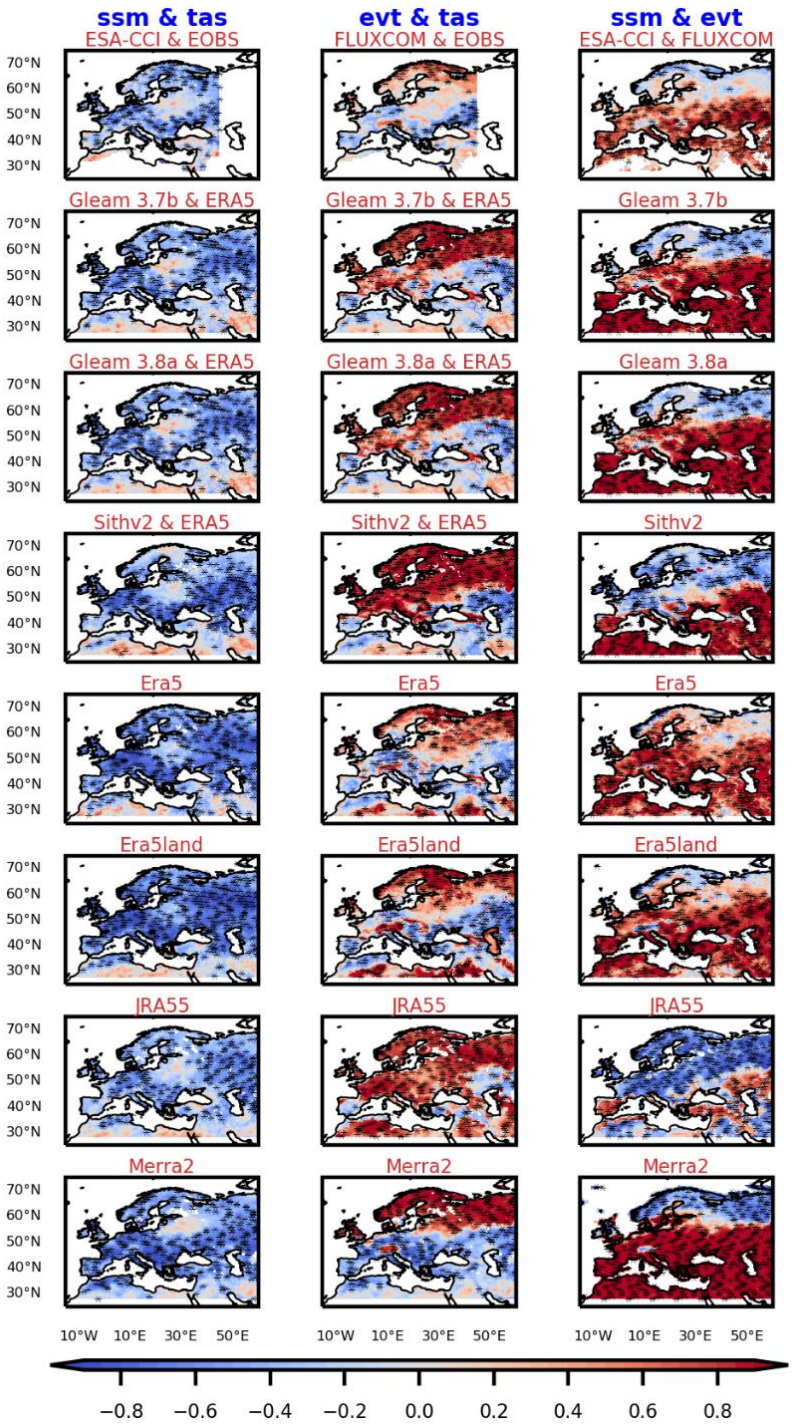

Figure 7: Correlation between surface soil moisture (SSM) and near-surface air temperature (TAS) (left), EVT and TAS (centre), and (right) SSM and EVT during boreal summer (JJA). The black dots represent correlations that are locally significant at the 95% significance level. The E-OBS, FLUXCOM, and ESA-CCI datasets were regridded to the common FLUXCOM grid at 0.5° resolution, while the SiTHv2, ERA5, and ERA5-Land datasets were regridded to the ERA5 grid at 0.25° resolution.

3.3.2 Process representation

Soil moisture (SM) anomalies play a critical role in sub-seasonal to seasonal forecasting by influencing surface energy and moisture exchange over land. But it is important that this coupled behaviour is well represented in forecasting systems. In this section we describe a number of ways of measuring soil moisture atmosphere coupling that will be used to assess model performance. These include correlation-based metrics, windows of opportunity, and multi-variate Machine Learning approaches.

3.3.2.1 Gridpoint correlation and co-variability metrics

Gridpoint correlations between seasonal means provide a measure of the strength of two aspects of this coupling on interannual timescales: the terrestrial leg (linking soil moisture and evapotranspiration) and the atmospheric leg (linking evapotranspiration and atmospheric variables, e.g. temperature, precipitation, cloud cover or circulation parameters). The chain of processes involved in land-atmosphere interactions that constitute the atmospheric-leg are complex (e.g Santanello et al, 2018). For example, soil moisture influences cloud amount and large-scale circulation, but the relationship is complex, and depends on environmental conditions, such as static stability (Huang and Margulis, 2011). Nevertheless, a comparison of the relationship between soil moisture and atmospheric variables captures the emergent behaviour resulting from these complex interactions and can be used to evaluate models.

Over Europe there is a large spatial variation in the correlation between evapotranspiration and near-surface air temperature: significantly positive over northern Europe and significantly negative over southern Europe. In contrast, the soil moisture–evapotranspiration relationship exhibits the opposite pattern. These differences reflect distinct regimes—northern Europe is energy-limited, where rising temperatures promote evaporation due to sufficient moisture availability, while southern Europe is water-limited, where evaporation is suppressed despite increasing temperatures due to limited soil moisture. However, the magnitude depends strongly on the datasets used to estimate this (see Figure 7).

A connected pathway from an initial soil moisture anomaly to atmospheric response can be identified and measured by looking at correlations between the variables involved in each step in the chain. We test whether dry anomalies at the start of the forecast (e.g. 1st July) are preferentially followed by anomalous temperature in month 2 (August). This can occur via the persistence of the 1 July soil moisture anomaly to 1 August, leading to low August evaporation and high August temperature. As such, this pathway can only occur in moisture-limited regions, where low soil moisture is associated with increased sensible heat flux. Although not all heatwaves are predictable in this way, dry July soil amplifies the risk of extreme August heat, with specific seasonal forecast ensemble members frequently following the pathway above in moisture-limited regions. We thus measure model fidelity by comparing the pathway steps in model hindcasts and observed data from GLEAM4 for the land surface (Miralles et al, (2025), and ERA5 for air temperature (Hersbach et al, (2020)). We make detrended area mean time series of ensemble seasonal model outputs for a set of 232 global land regions of roughly equal area (Stone, 2019) and then produce correlations for successive steps of the pathway. With only one 24-year realisation of hindcast-period observations, observed correlations are relatively uncertain, therefore we seek to assess the consistency of the model and observations given these uncertainties. We subsample the model ensemble, randomly picking one member per hindcast year to produce a single model realisation of the hindcast period. This makes 1000 such realisations, producing a distribution of model correlations, and we then estimate the percentile of observed correlations within this distribution. For the example of Texas (Figure 8), the observed soil moisture persistence and SM-E correlations look consistent with the model distribution, but the Evaporation-Temperature correlation suggests that the atmospheric leg of the coupling is too strong in the model. Model performance is scored as the number of regions where the observed percentile lies within the

model distribution for each pathway step. The new seasonal demonstrators will be assessed by comparing their score with the current C3S systems (phase zero demonstrators).

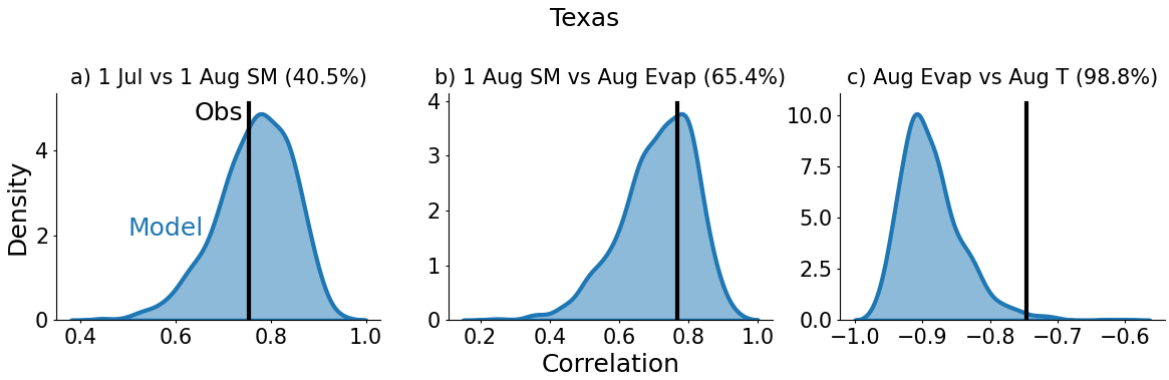

Figure 8: Subsampling tests of the successive links in the land surface pathway, for an example region. Panels show observed correlations (black lines) and probability density function of subsampled model correlations (for ECMWF-SEAS5) (shaded blue curves) for a) 1 July and 1 August soil moisture, b) 1 August soil moisture and mean August evaporation, c) mean August evaporation and mean August surface air temperature. The percentile of the observed correlation value within the model distribution is shown in each panel heading.

Variations on such metrics have also been investigated, such as the scaled correlation metrics defined in Dirmeyer (2011) and Dirmeyer et al. (2014). The first of these measures the strength of the terrestrial leg of the coupling (i.e. the link between the root-zone SM and evaporation (E):

$$I_{SM-E} = \sigma(E) \times \rho(SM, E)$$

In this metric the correlation between SM and E, ρ(SM,E), is scaled by the standard deviation of E, σ(E), to highlight regions where both the correlation and the climatological variance is high. The metric:

$I_{SM-X} = \sigma(X) \times \rho(SM, E) \times \rho(E, X)$,

where X is any atmospheric variable measures the combined strength of the terrestrial leg and the atmospheric leg. I_{SM-E} and I_{SM-X} have been applied to current C3S hindcasts (see Figure 9, taken from Day et al., 2025), and will be applied to new demonstrators.

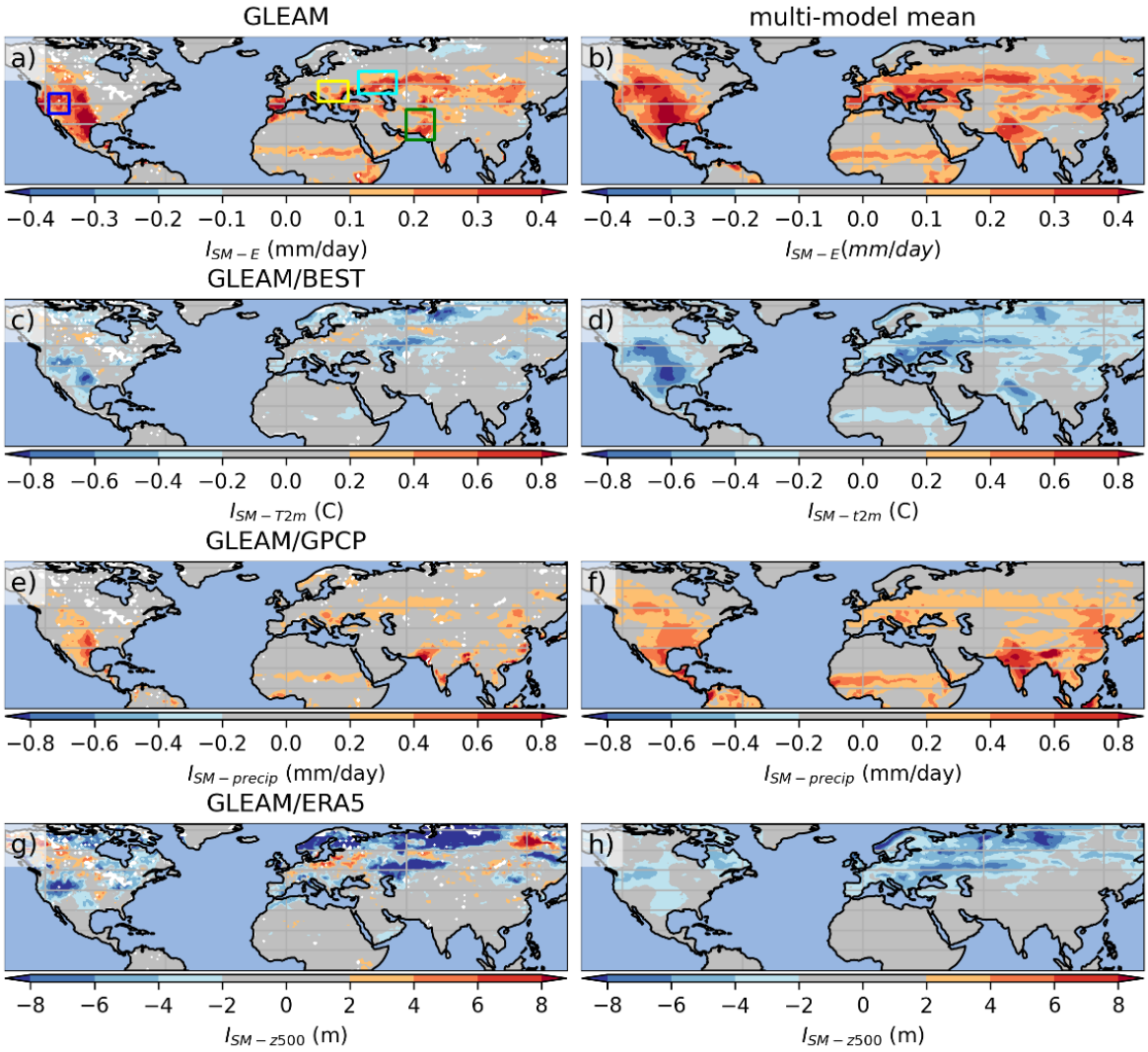

Figure 9: Application of soil moisture atmosphere coupling metrics to observations (BEST, GPCP) and reanalysis (GLEAM, ERA5) (left column) and C3S hindcasts (right column). The terrestrial leg metric ISM-E is shown in the top row, but the 2-legged metric for 2m-temperature, precipitation and z500 are shown below (figure from Day et al., 2025).

The correlation metrics above have been applied to monthly or seasonal mean data, but diagnostics on the sub-daily timescale captured by geostationary satellite data, such as that from MSG, have also been developed. The co-variability of observed LST and soil moisture is a useful reference to assess how well models represent surface processes. We have developed a novel metric, which will be used to compare reanalyses and seasonal forecasts with satellite data. It is based on the relationship between monthly SM and the monthly semi-diurnal amplitudes of LST (LST_Amp), which characterizes the monthly morning heating rate, which in turn relates to how efficiently available energy is used for evaporation. The metric is then computed as follows:

- For each month of the year, we compute the rank correlation between monthly surface soil moisture (SSM) and LST_Amp per grid-point;
- We then use the statistical significance of the rank correlation, p-value, to estimate the following metric:

$N_{covar} = \sum s(p_m)$, m= Jan, Feb, …, Dec

where s varies linearly between 1 (for p < 1%) and 0 (for p = 10% or higher).

N_{covar} measures the persistence of strong coupling between soil moisture and LST throughout the year, keeping in mind that the soil moisture will condition LST amplitudes (i.e., evapotranspiration) in water limited evaporative regimes.

In the example shown below (Figure 10), N_{covar} is derived from both ERA5 and satellite products. To ensure a fair comparison only cases with ERA5 total cloud cover below 30% are considered.

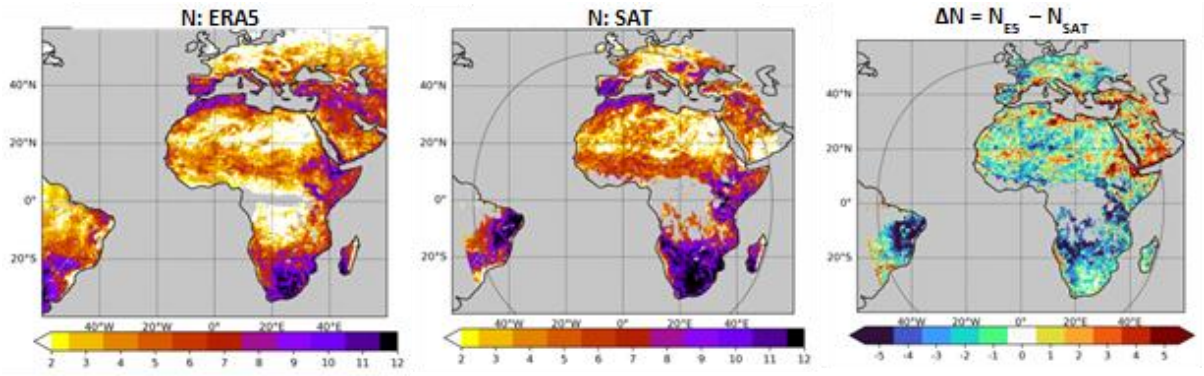

Figure 10: Co-variability metric N_{covar} between LST and SM, for ERA5 and for satellite observations (estimated for the 2004-2023 period). Pixels are shown in grey if all months have a sample size smaller than 10. (c): Difference of N_{covar} between ERA5 and satellite data. Grey mask in the ΔN map is the union of the masks in the two N_{covar} maps. The grey contour shows MSG viewing angles of 60º.

3.3.2.2 Spatial Covariances: Maximum Correlation Analysis

Maximum covariance analysis (MCA) looks for patterns in two space-time datasets which explain a maximum fraction of the covariance between them. It allows more spatially complex relationships between fields, such as the remote atmospheric response to forcing, to be captured, compared to the gridpoint correlation metrics described in 3.3.2.1. By introducing a lag between the fields used to perform the analysis it also has the potential to identify potentially predictable linkages between two variables.

In this case, 1m soil moisture, SM_{1m} at time t+$\boldsymbol{\tau}$ and Z_{200} at time t+$\boldsymbol{\tau}$ are expanded into K orthogonal signals:

$$Z_{200}(x,t) = \sum_{k=1}^{K} \mathbf{U}_k(x)\mathbf{a}_k(t)$$

$$SM_{1m}(x,t+\tau) = \sum_{k=1}^{K} \mathbf{V}_k\mathbf{b}_k(t+\tau)$$

where the covariance between $a_k(t)$ and $b_k(t+\boldsymbol{\tau})$ is the k^{th} singular value of the covariance matrix between SM and Z_{200}, decreasing for increasing k (see e.g. von Storch and Zwiers, 1999).

The patterns related to the leading mode between May soil moisture and JJA geopotential height at 200hPa (z200) are shown in Figure 11. It shows a z200 pattern which is similar to the so-called Circum-Global-Teleconnection described in Ding and Wang (2005) and indicates that it is preceded by dry soil moisture anomalies in the USA and Central Asia, which were

identified as regions of strong coupling using the gridpoint correlations. The technique will be used to reproducibility of patterns of large-scale variability in the seasonal hindcasts produced in CERISE.

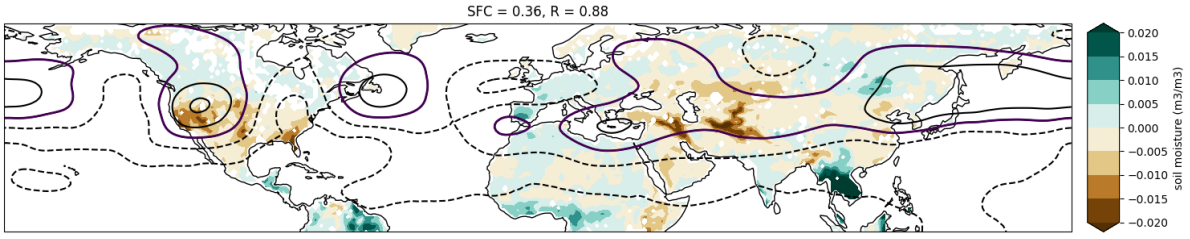

Figure 11: Patterns related to the leading coupled modes identified by the MCA analysis between May 1m soil moisture and JJA z200 field. The mode explains 36% of the combined variance between the two fields and the indices corresponding to these patterns have a correlation of 0.88.

3.3.2.3 Explainable AI to quantify driver importance

Understanding how current seasonal prediction systems model land-atmosphere interactions and subsequent temperature extremes is key for skillful seasonal predictions (Rind et al., 1982, Day et al., 2025). In this respect, a machine learning model is trained to forecast the likelihood/occurrence of daily temperature extremes at different locations, employing atmospheric and land surface variables as predictors (Figure 12a-b). The central concept is constructing the empirical model so that the individual contributions of the land surface and atmosphere can be easily traced and analyzed with Shapley Additive exPlanations (SHAP; Lundberg et al., 2017) values (Figure 12c). We leverage these different contributions in different regional case studies to better understand the roles of atmospheric and land drivers for temperature extremes at different locations, and how the different prediction systems represent such interactions. The analysis includes examining the atmospheric and land state configurations that are more prone to developing temperature extremes in reanalysis and prediction systems and detecting missing interactions and case study examples that might help explain poor predictability.

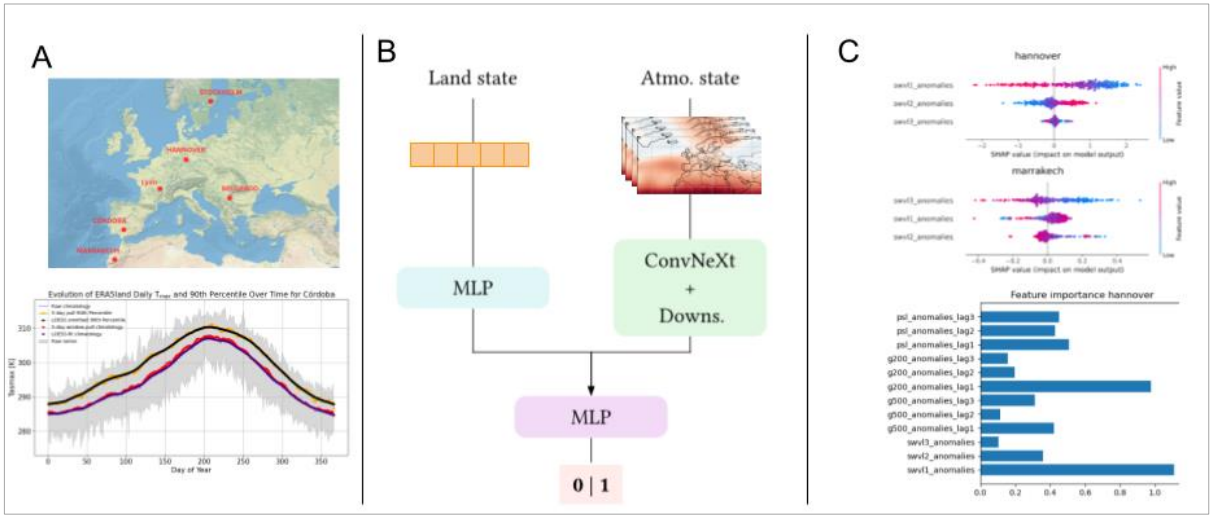

Figure 12: Scheme exemplifying the methodology. A: target temperature extreme definition and locations. B: deep-learning architecture composed of land and atmosphere state processing. C: explainability methods (examples in Hannover and Marrakech).

In building this tool, we employ the following input variables. The atmospheric circulation state is described by daily lags of geopotential height at 500 and 200 hPa (zg500 and zg200) plus sea level pressure (psl) anomalies. Lags from the preceding three days are used, while the spatial domain covers the Euro-Atlantic region (54W-70E and 14N-71N) at 1ºx1º resolution. The land state comprises local soil moisture (swvl) at the first three soil layers for the preceding week at the location of interest. For the study of reanalysis data, atmospheric variables from ERA5 are used, while ERA5-Land provides land variables. The target variable is the occurrence of local daily extreme temperature defined as the daily maximum temperature (tmax) exceeding the 90th percentile for the location of interest. To avoid high inter-day variability, we apply LOESS (*Mahlstein et al., 2015*) smoothing to the percentile climatology (see Figure 12a). For the study on reanalysis data, tmax from ERA5-Land is used.

The model architecture employs a Convnext (*Liu et al., 2022*) model for the atmospheric state and a multi-layer perceptron (MLP) to model the contributions from both the land state and the joint contribution of the land and atmosphere (Figure 12b). The GradientExplainer is used to compute the SHAP values, assessing the contributions from the different variables at different locations and case studies during the boreal summer. This data-driven model will be applied to evaluate the seasonal prediction systems with regard to the processes (atmospheric circulation and land-atmosphere interactions) driving heat extremes. In particular, we will assess whether and to what extent the variable importance of the different predictors is consistent between seasonal prediction models and observations, and whether the newly developed prediction demonstrators show improved representation compared to the current model versions.

3.4 Error growth in land-atmosphere coupling

3.4.1 Rationale

Biases in key variables, including surface temperature (T2m), temperature at 850 hPa (T850), mean sea-level pressure (MSLP), sensible and latent heat fluxes (SHF and LHF), and geopotential height (G500), significantly impact the predictive capability of seasonal forecasts initiated from November—a crucial period for establishing wintertime atmospheric patterns. The CMCC Seasonal Prediction System version 3.5 (SPS3.5) consistently exhibits notable negative biases in T2m, T850, and positive biases in MSLP, SHF, and LHF over Siberia. These biases indicate fundamental deficiencies in representing critical thermodynamic and dynamic processes. By comprehensively analyzing the evolution of these biases, we aim to elucidate reinforcing feedback mechanisms that could be targeted to improve model accuracy. The instrument developed and the mechanism studied then can be also used to investigate potentially similar behaviour in other seasonal prediction model simulations, such as those that are being generated in CERISE.

3.4.2 Methodological Framework

We developed a diagnostic framework centered on bias-spread diagrams (BSD) to systematically investigate bias evolution in different models as a starting framework to understand the possible key features to be analyzed, and to establish a protocol of investigation that could be used to test the improvements in different phases of simulation in WP5, entitled ‘Seasonal forecast demonstrators’.

3.4.2.1 Bias-Spread Diagrams (BSD)

The Bias-Spread Diagram (BSD) offers a visually succinct summary of models, enabling the identification of potential common mechanisms or divergent behaviors, particularly where specific variable biases grow anomalously compared to others. These diagrams are constructed by plotting two complementary metrics that evolve in time, producing a two-dimensional trajectory during the six months of the forecast. This allows us to observe how model biases develop and diverge throughout the forecast period, providing a powerful visualization for detecting emerging patterns and critical periods for model drift that could be a focus for improvement.

- For the x-axis, the spatially averaged bias is computed as the difference between each model and ERA5. This involves (i) computing the bias of each ensemble member relative to ERA5; (ii) calculating the ensemble mean bias; (iii) computing a spatial average over the analysed region; (iv) averaging for each day of the evolution across all years; (v) applying a 30-day running mean to isolate processes relevant to land-atmosphere interactions.
- For the y-axis, the metric represents the difference between the spatial standard deviations of the model and the one for ERA5. Similar to the bias calculation, this metric involves ensemble averaging, spatial and yearly averaging, and smoothing with a 30-day running mean to highlight variability patterns pertinent to diagnosing model behavior.

See Figure 13 for an example.

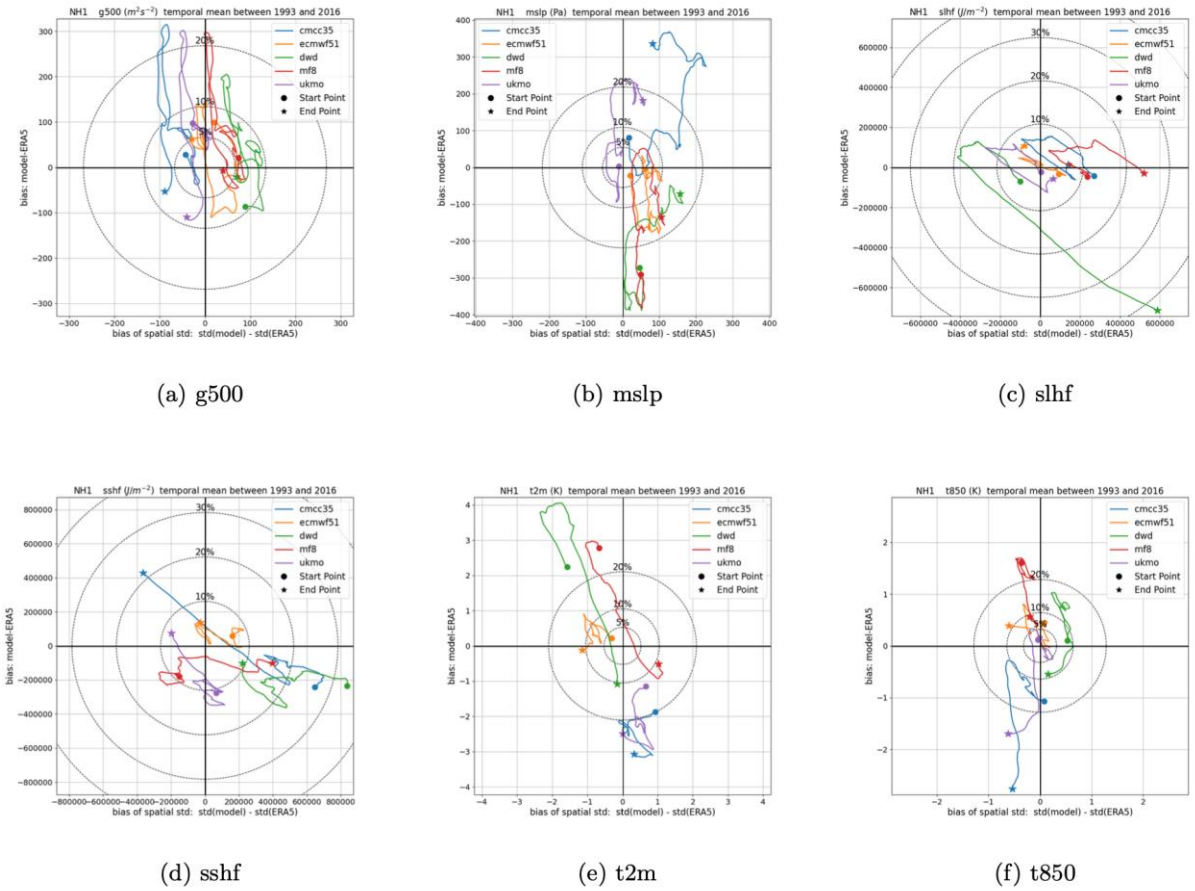

Figure 13: Bias-Spread Diagrams for key variables (MSLP, T850, T2m, SHF, LHF, G500) comparing CMCC SPS3.5, ECMWF SEAS5, DWD21, Meteo France System 8, and UK Met Office versions 600–602 (merged). Each plot shows the trajectory of the model's ensemble-mean bias (x-axis) and spatial standard deviation of the bias (y-axis) in the NH1 region (Siberia), smoothed with a 30-day running

mean. The trajectories span the period from day 15 to day 165 of the 180-day forecast, due to the application of the 30-day running mean. Circular dotted lines represent 25%, 50%, 75%, and 100% of the mean ERA5 standard deviation in the region.

3.4.2.2 Identification of Key Diagnostic Time Windows

From the previous analysis, for the CMCC model, three strategically selected periods were identified to systematically investigate different stages of bias evolution: an initial phase (Nov 20-Dec 10) representing the initial bias development, a transitional phase (Dec 30-Jan 19) corresponding to a period when snow cover is well-established over Siberia, and a later phase (Feb 8-28) following the onset of snow melt, aimed at analysing subsequent feedback reinforcement.

3.4.2.3 Local thermodynamic and large-scale dynamical mechanisms

Step 1: Investigate snow depth bias. Starting from the recognition of systematic snow depth biases in the CMCC model, we compare the model's snow depth fields against ERA5 to highlight regional anomalies, particularly during the transitional phase when snow cover is well-established.

Step 2: Analyse the relationship between snow and energy fluxes. To understand how snow anomalies influence surface-atmosphere energy exchanges, and check for potential model deficiency, for example in insulation properties, we compute the surface energy residual (E_r):

$$E_r = L_{net} + S_{net} - H - LH$$

This involves evaluating net shortwave (S_{net}) and longwave (L_{net}) radiation and subtracting the sensible (H) and latent heat (LH) heat fluxes. By comparing the energy residual in the model and ERA5, we assess whether snow biases lead to excessive insulation or energy trapping at the surface.

Step 3: Connect energy residuals to surface temperature biases. We assess the relationship between surface energy residual bias and 2-meter temperature (T2m) bias by analyzing their spatial co-variability. Specifically, we compute pointwise correlations between the T2m and energy residual biases across the domain and across ensemble hindcasts. To highlight statistically significant associations, we apply stippling to regions where these correlations exceed the 95% confidence level ($p < 0.05$). This approach reveals how spatial heterogeneity in surface energy exchanges is reflected in temperature biases, thereby identifying areas where land-atmosphere interactions play a dominant role in driving surface temperature errors.

Step 4: Assess vertical atmospheric structure. To understand possible decoupling between the surface and lower atmosphere, we examine vertical temperature profiles. Special attention is given to detecting the presence of inversion layers that could limit vertical heat exchange, especially during the transitional time window.

Step 5: Explore dynamical contributions. Finally, we explore the role of atmospheric circulation by examining MSLP patterns and associated wind anomalies. This helps determine whether dynamical processes, such as reinforced anticyclonic circulation, contribute to maintaining the cold bias over time. This also could explain why other regions at the same latitude with similar local response, and same parameterization for physical processes do not experience the same bias evolution.

This integrated diagnostic approach enables us not only to understand the feedback mechanisms contributing to bias growth in the CMCC model, but also to evaluate improvements in the future model demonstrators. Specifically, the same methodology can be applied to new simulations to assess whether: (i) the magnitude and temporal evolution of key biases (e.g., T2m, MSLP, T850) are reduced relative to ERA5; (ii) snow depth biases are mitigated, leading to smaller energy residual discrepancies and reduced impact on surface temperature; (iii) the vertical temperature structure shows improved coupling between surface and lower troposphere, limiting spurious inversions; and (iv) the strength and spatial extent of bias-induced circulation anomalies, particularly in MSLP, are diminished. These criteria provide an objective and transferable basis to quantify progress across model generations.

3.5 Consistency of hindcasts and forecasts

3.5.2 Significance of differences between hindcast and forecast initial conditions

To avoid large biases in seasonal forecasts of surface climate, it is important to ensure that the land surface is initialised consistently between forecasts and hindcasts. This arises as the hindcast is used to correct time-dependent biases in the forecast. Substantial temperature biases have previously been found to be artificially introduced when forecast and hindcast soil moisture initialisation were mismatched. We have built tools to check the consistency of initialised soil moisture and snow depth using area-mean time series for the IPCC WG1 reference regions (Iturbide et al, 2020). We obtain values for the first day of each member of hindcasts and forecasts initialised on the first day of each month. This is done for the full hindcast period and a year of forecasts. Systematic differences between the forecast and hindcast values may indicate inconsistency, but it may also be that the chosen forecast year happens to be an outlier by chance. To check that the hindcast and forecast initialisation are genuinely consistent, we construct synthetic random multiannual time series based on the autocorrelation, trend and annual cycle of the hindcast soil moisture and snow depth. These synthetic series cover both hindcast and forecast periods, so we can examine whether the initialised forecast data is consistent with a plausible extrapolation of the hindcast initial state. The metric for model performance is the number of regions where the initial forecast soil moisture or snow depth is significantly different from the synthetic series. This will allow us to compare the performance of seasonal hindcasts initialised using land-surface assimilation with those in the existing C3S database.

3.5.3 Comparison with operational analysis and offline Land Data Assimilation

Since reanalysis products like ERA5 or ERA5-Land, which are used to initialized seasonal re-forecasts, can be inconsistent with real-time initialization, some C3S re-forecasts use offline land surface reanalysis produced with the same land model version as the operational analysis. This is the case at ECMWF, where land surface fields are initialized from an open-loop land re-analysis. Offline re-analysis using land data assimilation instead of open-loop is expected to improve the consistency with the operational analysis. A methodology has been developed to determine if the CERISE land data assimilation (LDAS) re-analyses prototypes provide greater consistency with operational analysis than ERA5, ERA5-Land and open loop offline land re-analysis. This method consists of computing the root mean square error difference between the prototype land DA re-analysis (LDAS) and ERA5 (or ERA5LAND) using the ECMWF operational analysis as a reference. The formula is:

$$f(x) = \sqrt{\frac{\sum_{t=1}^{N}(\mathrm{REAN(x,t)} - OPER(x,t))^2}{N}} \; - \; \sqrt{\frac{\sum_{t=1}^{N}(\mathrm{OFFL(x,t)} - OPER(x,t))^2}{N}}$$

where REAN(x,t) is the land surface value (e.g. soil temperature level 1) of the reanalysis (ERA5 or ERA5-Land) at a given point (x) and time (t); OPER(x,t) is the corresponding value in the operational analysis and OFFL(x,t) is the value in the offline land DA reanalysis.

Positive values indicate a greater consistency with the operational analysis, and therefore an improvement (larger RMSD with ERA5/ERA5-Land than with the offline DA reanalysis) while negative values indicate a degradation. As an example, Figure 14 shows that the Volumetric soil water level 1 from the ERA6-Land pre-prototype is more consistent with the current ECMWF operational analysis than with ERA5 in many regions, although there are a few regions (blue color) where ERA5 is more consistent.

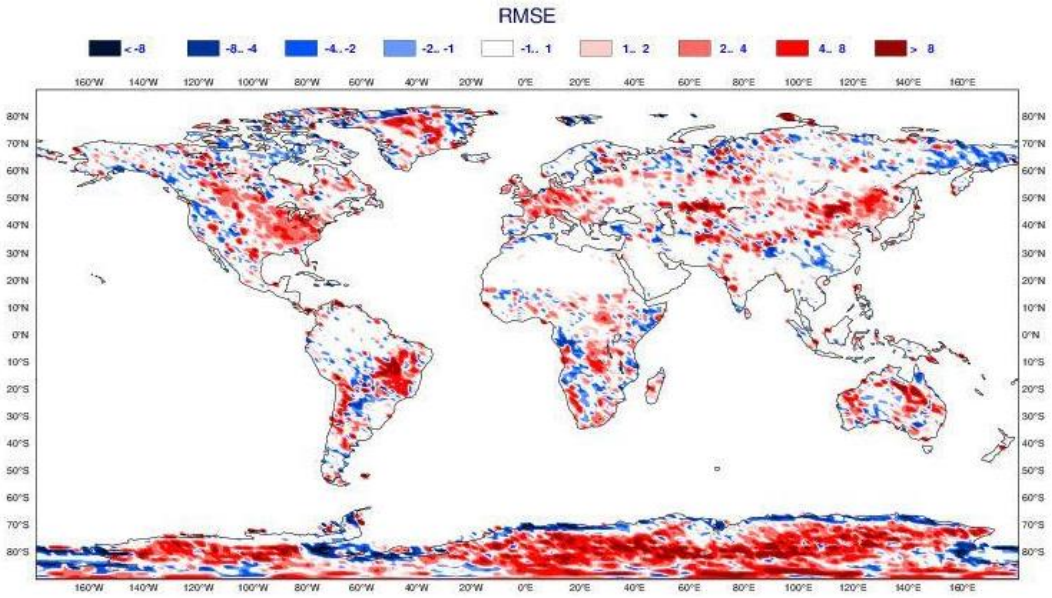

Figure 14: Impact of the ERA6-Land prototype on top layer soil moisture consistency shown as RMSE difference between ERA5 and ERA6-Land using the ECMWF operational analysis as a reference. Red (blue) colours indicate better (worst) consistency with the operational analysis for ERA6-Land than for ERA5.

3.6 Assessment of trends in land-surface variables in LDAS, ERA5 and ERA5-land

Trends in the land reanalysis used as initial conditions can significantly impact the representation of trends in the seasonal forecasts themselves. An error in the representation of seasonal trends can lead to significant errors in the model calibration and calculation of forecast anomalies. Therefore, it is important to initialize the seasonal forecasting systems with an analysis exhibiting realistic trends. The trends in the land surface variables produced by ERA6 pre-prototypes are compared with those produced by ERA5 and ERA5-Land using the following methodology: for each grid point and each soil variable, a linear regression is performed over the full time series (e.g. 1940-2020 for ERA5) to assess the amplitude of the trend. The linear regression coefficient is calculated using the formula:

$$b = \frac{n\left(\sum_{k=1}^{n} x*y\right) - \left(\sum_{k=1}^{n} x\right) * \left(\sum_{k=1}^{n} y\right)}{\left(n* \sum_{k=1}^{n} x^2\right) - \left(\sum_{k=1}^{n} x\right)^2}$$

where b is the trend coefficient, y the seasonal mean value of the land surface variable (e.g. soil temperature level 1) at a specific grid point and year (k). X(k) represents the k^{th} year from 1940 to 2020.

A statistical test is then performed by resampling the data using a bootstrap technique where 20 years are randomly removed from the time series and a linear regression is then performed on each of 10,000 new time series. If 99% of the 10,000 new time series display a trend with the same sign as the full time series, then the trend is considered to be statistically significant. Figure 15 shows an example for volumetric soil water level 1. Areas where trends are significantly different between ERA6-Land pre-prototypes and ERA5 or ERA5-Land are further investigated by analysing the time series and identifying possible discontinuities, which are an indication of improvements or degradation in the quality of ERA6-Land. This methodology will also be applied to diagnose 2-metre temperature trends in seasonal forecast demonstrators at various time ranges.

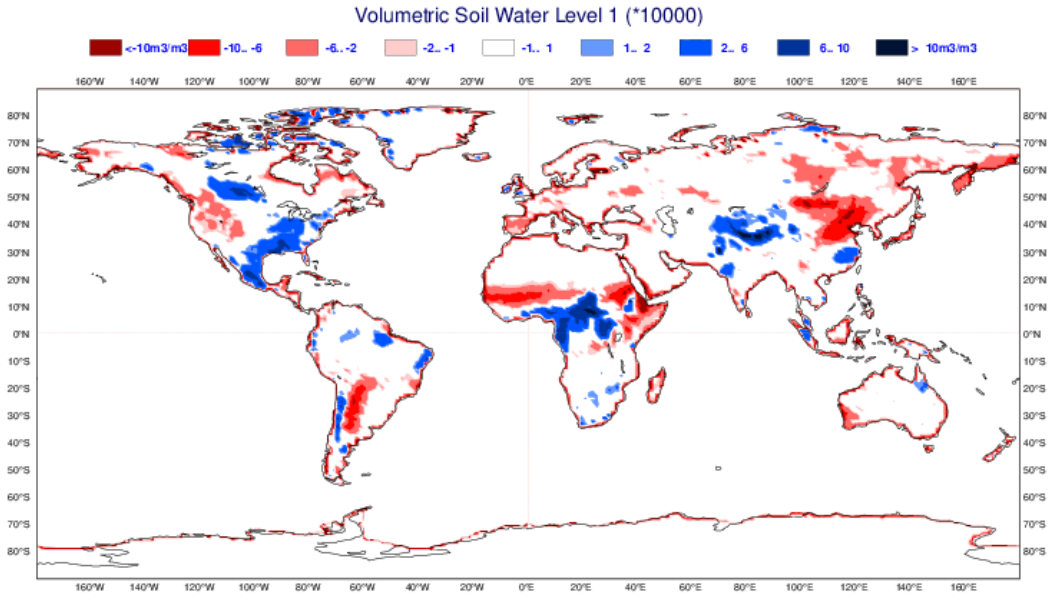

Figure 15: Trends in volumetric soil water level 1 in June-July-August computed from ERA5-Land over the period 1950-2020. Red (blue) colors indicate a negative (positive) trend. Areas where the trend is not statistically significant within the 1% level of confidence have been blanked.

4 Conclusion

During the first 30 months of the CERISE project, a variety of techniques and methodologies for evaluating the increased fidelity of land surface processes in reanalyses and seasonal forecast ensembles have been developed. These tools are intended to offer a wide-ranging capability for evaluating the new reanalysis prototypes and seasonal prediction demonstrators being produced in CERISE.

The range of diagnostic tools developed include methods to detect differences in skill and reliability of forecasts of snow cover, indicators for of snow phenomena such as snow onset and measures of snow-atmosphere coupling that account for the direction of causality. Methods to measure the verification of the spatial distribution of snow cover have also been developed. Hydrological simulation using reanalysis or seasonal forecast inputs provides a novel test of the fidelity of representation of land surface processes, with new diagnostic scores being produced to interpret the results.

Testing the quality of the representation of soil moisture and linked variables (e.g. evapotranspiration) will be essential for the new prototypes and demonstrators. The assessment of the available observations that we have conducted shows that observational uncertainty will need to be considered if meaningful conclusions are to be drawn from these tests. As for snow cover, methodologies for examining soil moisture's influence on the atmosphere have been developed. These examine links between soil moisture anomalies and subsequent warm season temperatures and other atmospheric variables. Further, even more sophisticated ways of testing soil-atmosphere coupling through satellite-derived diurnal cycle amplitudes have been produced, and statistical approaches to assess the non-local effects of soil water (through global patterns of atmospheric circulation) implemented. Machine learning approaches have also been harnessed to evaluate the relative roles of land surface influences compared to e.g. weather patterns in observations, reanalyses and seasonal modelling.

Methods for assessing the error growth in seasonal demonstrators have been produced. Analyses of these errors can reveal the physical causes and effects of errors in the initial land-surface conditions and in modelling. The test case here has errors in the representation of Siberian snow cover. Examination of errors in the new demonstrators will reveal whether they are behaving differently from early phases without land surface initialisation.

Consistency of forecast and hindcast land surface initialisation is key for the production of unbiased forecasts, and methods to assess the geographical extent of this consistency in new demonstrators have been produced as part of this work. Further to this, comparison methods have also been developed to assess whether land-surface initialisation values are closer to offline data assimilation than other datasets. Finally, tools to evaluate trends in land-surface variables in reanalyses and seasonal ensembles have been created to assess whether trends become more accurate in new experimental products.

The new toolkit described in this report aims to provide information on the fidelity of processes related to the representation of the land surface that supplements information from the standard verification (using skill scores) that will be undertaken in WP5. These standard methods can struggle to identify improvements, as small changes can require very large data samples in order to be statistically unambiguous. By using a range of novel additional approaches to assess the quality in the new reanalyses and the seasonal ensembles, we hope to improve our chances of gathering evidence of improvements resulting from the assimilation of land surface data. Despite this, the approaches we have developed are experimental at this stage, and are not guaranteed to be successful in providing clear evidence in every case. Nevertheless, our strategy of employing a wide range of variables, datasets and techniques in our assessment protocol is designed to maximise the chances of making a clear determination on the benefits of the innovations under trial in CERISE. The most relevant tools will be identified later in the project with recommendation for potential usage for operational assessment of future C3S reanalyses and seasonal prediction systems.

5 Bibliography

Adler RF, Sapiano MRP, Huffman GJ, Wang J-J, Gu G, Bolvin D, Chiu L, Schneider U, Becker A, Nelkin E, Xie P, Ferraro R, Shin D-B (2018) The Global Precipitation Climatology Project (GPCP) Monthly Analysis (New Version 2.3) and a Review of 2017 Global Precipitation. Atmosphere 9(4):138. https://doi.org/10.3390/atmos9040138

Copernicus Climate Change Service (C3S), Climate Data Store (CDS), (2022): Cloud properties global gridded monthly and daily data from 1982 to present derived from satellite observations. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). DOI: 10.24381/cds.68653055 (Accessed on 02-APR-2025)

Day J., Vitart, F., Stockdale, T. et al. Soil-moisture-atmosphere coupling hotspots and their representation in seasonal forecasts of boreal summer, 08 May 2025, PREPRINT (Version 2) available at Research Square [https://doi.org/10.21203/rs.3.rs-5483979/v2]

Ding, Q., and B. Wang, 2005: Circumglobal Teleconnection in the Northern Hemisphere Summer. *J. Climate*, **18**, 3483–3505, https://doi.org/10.1175/JCLI3473.1.

Dirmeyer PA (2011) The terrestrial segment of soil moisture–climate coupling. Geophysical Research Letters 38(16). https://doi.org/10.1029/2011GL048268

Dirmeyer PA, Wang Z, Mbuh MJ, Norton HE (2014) Intensified land surface control on boundary layer growth in a changing climate. Geophysical Research Letters 41(4):1290–1294. https://doi.org/10.1002/2013GL058826

Goessling HF, Jung T. A probabilistic verification score for contours: Methodology and application to Arctic ice-edge forecasts. Q J R Meteorol Soc. 2018; 144:735–743. https://doi.org/10.1002/qj.3242

Henck, A. C., Huntington, K. W., Stone, J. O., Montgomery, D. R., and Hallet, B., 2011: Spatial controls on erosion in the Three Rivers Region, southeastern Tibet and southwestern China, Earth and Planetary Science Letters, 303, 71–83, https://doi.org/10.1016/j.epsl.2010.12.038.

Hersbach H, Bell B, Berrisford P, et al. The ERA5 global reanalysis. Q J R Meteorol Soc. 2020; 146: 1999–2049. https://doi.org/10.1002/qj.3803

Huang, H., and S. A. Margulis, 2011: Investigating the Impact of Soil Moisture and Atmospheric Stability on Cloud Development and Distribution Using a Coupled Large-Eddy Simulation and Land Surface Model. J. Hydrometeor., 12, 787–804, https://doi.org/10.1175/2011JHM1315.1.

Iturbide, M. et al., 2020: An update of IPCC climate reference regions for subcontinental analysis of climate model data: definition and aggregated datasets. Earth System Science Data, 12(4), 2959–2970, doi: 10.5194/essd-12-2959-2020.

Kobayashi, Shinya, Yukinari Ota, Yayoi Harada, Ayataka Ebita, Masami Moriya, Hirokatsu Onoda, Kazutoshi Onogi et al. "The JRA-55 reanalysis: General specifications and basic characteristics." Journal of the Meteorological Society of Japan. Ser. II 93, no. 1 (2015): 5-48.

Komatsu, K.K., Takaya, Y., Toyoda, T. and Hasumi, H., 2023. A submonthly scale causal relation between snow cover and surface air temperature over the autumnal Eurasian continent. Journal of Climate, 36(15), pp.4863-4877.

Koster, R.D., Sud, Y.C., Guo, Z., Dirmeyer, P.A., Bonan, G., Oleson, K.W., Chan, E., Verseghy, D., Cox, P., Davies, H. and Kowalczyk, E., 2006. GLACE: the global land–atmosphere coupling experiment. Part I: overview. Journal of Hydrometeorology, 7(4), pp.590-610.

Lundberg, S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In arXiv [cs.AI]. http://arxiv.org/abs/1705.07874

Li, F., Orsolini, Y.J., Keenlyside, N., Shen, M.L., Counillon, F. and Wang, Y.G., 2019. Impact of snow initialization in subseasonal-to-seasonal winter forecasts with the Norwegian Climate Prediction Model. Journal of Geophysical Research: Atmospheres, 124(17-18), pp.10033-10048.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A ConvNet for the 2020s. In arXiv [cs.CV]. http://arxiv.org/abs/2201.03545

Mahlstein, I., Spirig, C., Liniger, M. A., & Appenzeller, C. (2015). Estimating daily climatologies for climate indices derived from climate model data and observations. Journal of Geophysical Research Atmospheres, 120(7), 2808–2818. https://doi.org/10.1002/2014JD022327

Miralles, D.G., Holmes, T.R.H., de Jeu, R.A.M., Gash, J.H., Meesters, A.G.C.A., Dolman, A.J. Global land-surface evaporation estimated from satellite-based observations, Hydrology and Earth System Sciences, 15, 453–469, doi: 10.5194/hess-15-453-2011, 2011

Miralles, D.G., Bonte, O., Koppa, A. et al. GLEAM4: global land evaporation and soil moisture dataset at 0.1° resolution from 1980 to near present. Sci Data 12, 416 (2025). https://doi.org/10.1038/s41597-025-04610-y

Munier, S. and Decharme, B.: River network and hydro-geomorphological parameters at 1/12◦ resolution for global hydrological and climate studies, Earth System Science Data, 14, 2239–2258, https://doi.org/10.5194/essd-14-2239-2022, 2022.

Muñoz-Sabater J, Dutra E, Agustí-Panareda A, Albergel C, Arduini G, Balsamo G, Boussetta S, Choulga M, Harrigan S, Hersbach H, Martens B, Miralles DG, Piles M, Rodríguez-Fernández NJ, Zsoter E, Buontempo C, Thépaut J-N (2021) ERA5-Land: a state-of-the-art global reanalysis dataset for land applications. Earth System Science Data 13(9):4349–4383. https://doi.org/10.5194/essd-13-4349-2021

Narváez-Campo, G. and Ardilouze, C.: Skilful Seasonal Streamflow Forecasting Using a Fully Coupled Global Climate Model, EGUsphere [preprint], https://doi.org/10.5194/egusphere-2024-2962, 2024.

Niu, G.Y. and Yang, Z.L., 2007. An observation-based formulation of snow cover fraction and its evaluation over large North American river basins. Journal of geophysical research: Atmospheres, 112(D21).

Rind, D. (1982). The influence of ground moisture conditions in north America on summer climate as modeled in the GISS GCM. Monthly Weather Review, 110(10), 1487–1494. https://doi.org/10.1175/1520-0493(1982)110<1487:tiogmc>2.0.co;2

Roberts, N. M., and H. W. Lean, 2008: Scale-Selective Verification of Rainfall Accumulations from High-Resolution Forecasts of Convective Events. Mon. Wea. Rev., 136, 78–97, https://doi.org/10.1175/2007MWR2123.1

Rohde R, Muller R, Jacobsen R, Perlmutter S, Mosher S (2013) Berkeley Earth Temperature Averaging Process. Geoinfor Geostat: An Overview 01(02). https://doi.org/10.4172/2327-4581.1000103

Stone, D. A. A hierarchical collection of political/economic regions for analysis of climate extremes. Clim. Change 155, 639–656 (2019).

von Storch and Zwiers (1999) Canonical Correlation Analysis, in Statistical Analysis in Climate Research.

Takaya, Y., Komatsu, K.K., Ganeshi, N.G., Toyoda, T. and Hasumi, H., 2024. A sub-monthly timescale causality between snow cover and surface air temperature in the Northern Hemisphere inferred by Liang–Kleeman information flow analysis. Climate Dynamics, 62(4), pp.2735-2753.

Weisheimer, A. and Palmer, T.N., 2014. On the reliability of seasonal climate forecasts. Journal of the Royal Society Interface, 11(96), p.20131162.

Xu, L. and Dirmeyer, P., 2011. Snow-atmosphere coupling strength in a global atmospheric model. Geophysical Research Letters, 38(13).

Zhang, K., Chen, H., Ma, N., Shang, S., Wang, Y., Xu, Q., & Zhu, G. (2024). A global dataset of terrestrial evapotranspiration and soil moisture dynamics from 1982 to 2020. Scientific Data, 11(1), 445.

Document History

Version	Author(s)	Date	Changes
0.1	Markus Donat (BSC), Jeff Knight (MetO), Frederic Vitart (ECMWF) plus WP6 partners	10th June 2025	Initial version
1.0	Markus Donat (BSC), Jeff Knight (MetO), Frederic Vitart (ECMWF) plus WP6 partners	26th June 2025	Final version after internal review

Internal Review History

Internal Reviewers	Date	Comments
Hauke Schulz (DMI) and Filipe Aires (ESTELLUS)	June 2025	Initial version V0.1
Patricia de Rosnay	25 June 2025	Comments on V0.1