EMI R&D PROJECT PROGRESS REPORT

All the following mandatory information needs to be provided. The length should *reflect the complexity and duration* of the project.

Reporting year	2025-2026
Project Title:	Towards a land/atmosphere/ocean/ice coupled data assimilation system: 20-year coupled reanalysis to initialise longer time-range forecasts, specifically seasonal forecasts (towards C3S multi-model project)
Computer Project Account:	Spitcard
Principal Investigator(s):	Denis Eiras, Marcelo Guatura, Giovanni Conti, Carla Cardinali
Affiliation:	CMCC
Name of ECMWF scientist(s) collaborating to the project (if applicable)	NA
Start date of the project:	April 2025
Expected end date:	April 2028

Computer resources allocated/used for the current year and the previous one

(if applicable)

Please answer for all project resources

		Previous year		Current year	
		Allocated	Used	Allocated	Used
High Performance Computing Facility	(units)	300000000	3322583	220000000	148514045
Data storage capacity	(Gbytes)	100000	14000	100000	67940

Summary of project objectives (10 lines max)

The project aims to develop a coupled data assimilation framework for the land, atmosphere, ocean and sea-ice components of the Earth system. The goal is to improve the consistency of initial conditions used for seasonal and longer-range forecasts. During the project, we investigate both traditional and machine-learning-based approaches for coupled data assimilation, with particular focus on computational efficiency and scalability. The work is organised in successive stages, starting from land–atmosphere coupling, followed by the inclusion of the ocean component. A final objective is to produce a coupled reanalysis prototype, initially for a limited period as a proof of concept, before considering longer reanalysis production.

Summary of problems encountered (10 lines max)

The original project relied on the SPREADS coupled data assimilation framework to generate a long coupled reanalysis. Initial experiments showed that extending this approach to a fully coupled multi-component reanalysis would require significantly larger computational resources than originally anticipated and available. This motivated the development of an alternative machine-learning-based coupled data assimilation framework able to retain cross-component information while reducing computational cost. During the reporting period, the main effort was therefore devoted to the design, implementation and testing of this new framework and to demonstrating its ability to reproduce coupled land–atmosphere analyses.

Summary of plans for the continuation of the project (10 lines max)

The next phase of the project will extend the current proof-of-concept system from a European domain to a global configuration. The machine-learning framework will be expanded from boundary-layer atmospheric variables to the atmospheric column that is most affected by the land processes. In this next stage, the present ERA5-Land and ERA5 fields used in the proof-of-concept experiments will be replaced by SPREADS-Land and SPREADS-Atm, in order to build the coupled framework on the same data assimilation systems developed and operational at CMCC. The observing system will also be enlarged, with a much stronger use of satellite observations together with additional land observations, including flux-tower measurements and other surface datasets relevant for land–atmosphere interactions. The evaluation of the new experiments will be based primarily on independent observations. In the final phase, the ocean component will be incorporated to move towards a fully coupled land–atmosphere–ocean data assimilation framework

List of publications/reports from the project with complete references

Cardinali, C., Conti, G., Guatura, M., Saarinen, S., Gonçalves De Gonçalves, L., G., Anderson, J., and Reader, K., Eiras, D., 2026. SPREADS: From Research to Operational Open-Source Data Assimilation System. Submitted to GMD

de Gonçalves, L. G. G., de Ávila, Á. V., Cardinali, C., Conti, G., Guatura, M., Peano, D., Raczka, B., Benassi, M., Eiras, D., Gualdi, S., 2026. CMCC-LDAS: A 21-year global multivariate ensemble land data assimilation reanalysis 2002–2022. Submitted to Advances in Modeling Earth Systems.

Guatura M., Conti G., Gonçalves De Gonçalves L. G., Cardinali C., Vasconcelos A., Gualdi S., Eiras D., 2026. A Unified Graph Network Framework for Coupled Forecasting and Data Assimilation: Application to Soil Moisture Dynamics. In preparation.

Symposium ISDA - Nanjing, China 31/08 - 04/09 https://da2026.leqiaoxueshu.com
A 21-Year Multi-Stream Land Reanalysis (2002–2022) Using the CMCC-LDAS: Simultaneous Assimilation of Soil, Snow, and Vegetation
Luis Gustavo Gonçalves de Gonçalves[1], Carla Cardinali[1], Alvaro Vasconcellos[1], Giovanni Conti[1], Marcelo Guatura[1], Brett Raczka[2], Silvio Gualdi[1]
1 Centro Euro-Mediterraneo sui Cambiamenti Climatici (CMCC)
2 National Center for Atmospheric Research (NCAR)

Summary of results

If submitted **during the first project year**, please summarise the results achieved during the period from the project start to June of the current year. A few paragraphs might be sufficient. If submitted **during the second project year**, this summary should be more detailed and cover the period from the project start. The length, at most 8 pages, should reflect the complexity of the project. Alternatively, it could be replaced by a short summary plus an existing scientific report on the project attached to this document. If submitted **during the third project year**, please summarise the results achieved during the period from July of the previous year to June of the current year. A few paragraphs might be sufficient.

Development of the GAT-Land-Atm framework

Introduction

The original objective of the project was to investigate coupled data assimilation approaches for land, atmosphere, ocean and sea-ice applications using the **SPREADS** framework. During the first experiments it became clear that extending a traditional ensemble-based coupled assimilation system towards a long coupled reanalysis would require very large computational resources. For this reason, part of the project effort was redirected towards the development of a **machine-learning-based coupled data assimilation framework**.

The framework developed during this reporting period is called **GAT-Land-Atm** and focuses on the coupled **land–atmosphere** problem. **GAT-Land-Atm** is based on **GAT-Land** developed by Guatura *et al*., 2026. The objective is not to replace SPREADS, but to learn from the land and atmospheric backgrounds together with the observations how information is transferred between the land and atmosphere during the analysis step. Once trained, the machine-learning framework can generate coupled analyses at a fraction of the computational cost required by a traditional ensemble assimilation system.

The methodology is based on **graph neural networks**. In this approach, the model domain is represented as a graph in which each node corresponds to a grid point and neighbouring nodes are connected so that information can be exchanged spatially. This is particularly suitable for geophysical applications because land and atmospheric conditions at one location are influenced by the surrounding state through transport, surface fluxes and boundary-layer processes. The graph structure therefore provides a natural way to represent these spatial dependencies within a coupled assimilation framework.

From an architectural point of view, **GAT-Land-Atm** is organised in **two phases**. **PHASE0** is the land-background generation step. Its role is to produce the land background field, which in the present proof-of-concept configuration is the **surface soil-moisture background**. PHASE0 uses a temporal graph-convolutional model that learns the evolution of soil moisture from its previous states and provides the land background required by the coupled assimilation system.

PHASE1 is the actual coupled assimilation step of the framework. In this phase, the model receives as input the **land background from PHASE0**, the **atmospheric background from ERA5**, and the available **land and atmospheric observations**. These inputs are processed by a graph-attention coupler that allows each grid point to exchange information with neighbouring nodes. In this way, the correction applied at one location can depend not only on local observations, but also on the surrounding land and atmospheric state. **PHASE1** then produces two outputs: a corrected **land analysis** and a corrected **atmospheric analysis**. The atmospheric analysis is obtained by correcting the ERA5 atmospheric background, while the land analysis is obtained by correcting the PHASE0 land background. Through this design, information from land observations can influence atmospheric variables, and atmospheric observations can in turn contribute to the correction of the land state.

A key element of the PHASE1 architecture is that the model does not predict absolute analysed land and atmospheric states directly. Instead, it learns **bounded residual increments** that are applied to the background fields. In practice, the network predicts the correction that must be added to the PHASE0 land background and to the ERA5 atmospheric background in order to produce the final coupled analyses. This residual formulation is close to the logic of a classical data assimilation system, where the analysis is obtained by updating a background state using observations. In the present machine-learning framework, the corrections are learned from the training data using a dual loss function that combines an observation-space loss, computed from the mismatch with the available observations, with a background regularisation term that penalises the departures from the background state. This second term acts as a Tikhonov regulariser, helping to stabilise the inherently ill-posed inverse problem while ensuring that the analysis remains physically consistent with the background information.

The proof-of-concept system was intentionally kept simple and focused on the variables most relevant for **land–atmosphere coupling**. For the **land component**, the main analysed variable is **surface soil moisture**. For the **atmospheric component**, the first experiments concentrated on **near-surface and lower-atmosphere variables**, including **precipitation, 2-metre temperature, near-surface humidity, 10-metre winds and mean sea-level pressure**. These variables were selected because they are directly linked to land-surface conditions and boundary-layer processes and therefore provide a suitable first test of coupled land–atmosphere assimilation.

A key advantage of the GAT-Land-Atm approach is that it avoids the explicit estimation of **land–atmosphere error covariances** from large ensembles, which is one of the main computational bottlenecks of classical coupled data assimilation. Instead, the relationships between land and atmospheric variables are learned directly from the training data and from the observational constraints. This makes the framework computationally attractive and provides a practical basis for future extensions to a **global domain**, a **larger observing system**, and, in a later stage of the project, the inclusion of the **ocean component**.

Experimental Setup

The experiments carried out during the reporting period were designed as a proof of concept for a coupled land–atmosphere data assimilation framework over Europe. The objective was to test whether a machine-learning system can use land and atmospheric observations together with background information to produce improved analyses for both components within a common coupled framework.

The **GAT-Land-Atm** system is organised in **two phases. PHASE0** is the land-background learning step. In this phase, the model is trained using **ERA5-Land surface soil-moisture fields** over the period **2008–2011**. The purpose of PHASE0 is to learn how to generate the land background information (here defined as **BG**) required by the coupled framework. After training, the PHASE0 model is applied over the period **2012–2017**, and its performance is monitored through the corresponding loss function in order to verify that the learned representation remains stable over a period that is independent from the training years. In this way, PHASE0 provides the land background information (BG) used by the coupled assimilation step.

PHASE1 is the coupled land–atmosphere assimilation step. In this phase, the model combines three sources of information:

(i) **the land background fields produced by PHASE0**,
(ii) **the ERA5 atmospheric state**, and
(iii) **the available land and atmospheric observations**.

Using these inputs, PHASE1 produces corrected land analyses and corrected atmospheric analyses. The **training** of **PHASE1** was carried out over the period **2012–2015**.

After this training stage, the **coupled PHASE1** system was run in **cycling mode** over the period **2016–2017**. This two-year period is **the scientific evaluation period of the present report**. During this phase, the coupled assimilation framework was applied at 6-hourly frequency (00, 06, 12 and 18 UTC) over the European domain, using the observations defined for the experiment and producing the land and atmospheric analyses at each analysis time. The scientific results discussed in this report therefore refer to the analyses generated by the PHASE1 coupled system during **2016–2017, which is named GAT_AN**.

The proof-of-concept configuration focuses on variables directly linked to the land–atmosphere interactions. For **the land component**, the analysed variable is surface **soil moisture**. For the **atmospheric component**, the framework predicts **10-m wind components** (u10m and v10m), **mean sea-level pressure**, **2-m temperature**, **precipitation**, **incoming short-wave radiation** and **near-surface specific humidity**. These variables were selected because they are strongly influenced by land-surface conditions and boundary-layer processes and therefore provide a suitable first test of the coupled framework.

The experiments were performed over **Europe** using a **1° graph backbone** and a **6-hourly analysis cycle**. The observational constraints were chosen to represent both land and atmosphere. For the land component, the framework uses ESA-CCI soil-moisture satellite derived product. For the atmospheric component, it uses a combination of conventional SYNOP observations and IASI satellite product, with particular relevance for temperature- and humidity-related diagnostics. The coupled assimilation model is formulated and evaluated directly in **observation space**, meaning that the machine-learning corrections are constrained using the mismatch between the analysed fields and observations at the observation locations.

The scientific evaluation of the coupled analyses was performed over the 2016–2017 cycling period. The land and atmospheric analysis fields produced by PHASE1 were compared with **independent observations of the same type as those used in the assimilation, but not assimilated by the system**. For the land component, the machine-learning soil-moisture analyses were evaluated against **withheld** ESA-CCI soil-moisture derived observations. For the atmospheric

component, the analysed atmospheric fields were evaluated against **withheld** SYNOP and IASI observations.
Moreover, three PHASE1 configurations are tested, which differ mainly in how strongly the atmospheric background is corrected and in how much importance is given to atmospheric observations during training.
The L**and-Optimal configuration (LO)** applies a more conservative atmospheric correction and does not use observation-error weighting in the loss function. It also uses a simpler network configuration. In practice, this setup favours the **land analysis**, especially soil moisture, while producing only a weak atmospheric correction.
The B**alanced configuration (Bal)** applies a somewhat stronger atmospheric correction and gives more weight to the atmospheric observations during training, while still keeping a constraint towards the atmospheric background. In practice, this configuration preserves a clear positive impact on the land analysis but also allows a stronger atmospheric correction than **LO**, giving the best compromise between land and atmosphere.
The A**tmosphere-Optimal configuration (AO)** gives a much stronger weight to atmospheric observations in the training loss. This pushes the model to fit the atmospheric observations more closely and therefore maximises the atmospheric skill, especially for **precipitation**, **near-surface humidity** and **temperature-sensitive diagnostics**, but with a small reduction of the land skill.

Results

The validation of the GAT-Land-Atm framework was performed over the **2016–2017 coupled cycling period** by comparing the land and atmospheric analyses produced by **PHASE1** with **independent observations of the same type as those used in the assimilation, but not assimilated by the system**. The evaluation was carried out in **observation space**. For the **land component**, the reference field is the **ERA5-Land**. For the **atmospheric component**, the reference field is the **ERA5 atmospheric state**, while the corrected atmospheric field produced by PHASE1 is denoted here as **GAT_AN**. The validation was based on **RMSE**, **Bias**, **Skill Score**, **Taylor diagrams**, and **spatial ΔRMSE maps**.

Root-mean-square error and skill score

The first clear result of the validation is that the coupled framework improves the **surface soil-moisture analysis** in all tested configurations. The background RMSE (**BG versus observations**) for land soil moisture is **0.136,** which quantify the improvement with respect to the observations of the PHASE0 output**. GAT_AN** (PHASE1 coupled analysis output) reduces this error in all three configurations accordingly to the skill-score ($SS= 1-RMSE(GAT\_AN)^2 / RMSE(ERA5_{Atm}^{Land})^2$). The strongest land result is obtained in the **LO configuration**, for which the soil-moisture skill score reaches **0.518** (Table 1). This corresponds to a substantial error reduction and confirms that the coupled framework is able to recover a large part of the useful land correction. The **Bal configuration** shows a strong improvement for soil-moisture with a skill of **0.304;** the **AO configuration** still provides a small but meaningful improvement **0.106**. Table 1 below summarise the SS results.

For the **atmospheric component** the validation shows that **10-m wind components** and **mean sea-level pressure** have a small error reduction: it is important to note that the ERA5 field used to initialise PHASE1 is already of very good quality. In fact, their normalised RMSE is approximately **0.027** for **u10m**, **0.026** for **v10m**, and **0.011** for **MSLP**. In these conditions there is limited room for improvement, and the corresponding skill scores remain close to zero in all tested configurations. A similar behaviour is found for **screen-level temperature** from SYNOP observations.

A different behaviour is found for the variables that still have clear room for improvement and are more directly connected to land–atmosphere interactions. In particular, **precipitation**, **temperature**

(IASI derived profile), and **near-surface humidity** show the most encouraging atmospheric response. Their standardised RMSE values are larger than those of wind and pressure, with values of **0.189** for **precipitation**, **0.030** for **IASI T2m**, and **0.049** for **near-surface humidity**. These are also the variables for which the coupled framework produces the clearest positive atmospheric skill. In **Bal**, the skill scores reach **0.121** for precipitation, **0.217** for IASI T2m, and **0.208** for near-surface humidity (see Table 1). In the **AO configuration**, the corresponding values increase further to **0.134**, **0.255**, and **0.214**, respectively. These results indicate that the coupled analysis is able to transfer useful information from land and atmospheric observations into the atmospheric state, particularly for the variables most sensitive to land-surface conditions and boundary-layer processes.

The **screen-level temperature** and **MSLP** results remain positive but small in the balanced and atmosphere-oriented configurations. For example, the balanced configuration gives a skill of **0.028** for SYNOP T2m and **0.009** for MSLP, while the atmosphere-optimal configuration gives **0.026** and **0.014**, respectively. These values are modest, but they remain consistent with the fact that the ERA5 provided background for these variables is already very accurate in the present proof-of-concept setup.

Bias

Bias was used as a complementary diagnostic to verify whether the machine-learning correction introduced systematic shifts relative to the observations. The validation indicates that the coupled analysis does **not** generate large systematic biases in the corrected fields. The main benefit of the coupled framework is therefore not a simple mean adjustment, but a reduction of the **random and structured component of the error**, especially for **soil moisture**, **precipitation**, **IASI T2m**, and **near-surface humidity**. This is an important result because it suggests that the coupled correction is improving the agreement with the observed variability rather than only shifting the mean state.

Taylor-diagram diagnostics

The Taylor diagrams provide a compact summary of the relationship between the analysed fields and the observations in terms of **correlation**, **normalised standard deviation**, and **distance from the observational reference**. The diagrams confirm the contrast already seen in the RMSE and skill-score diagnostics.

For **u10m**, **v10m**, **MSLP**, and to a lesser extent **SYNOP T2m**, the analysis points remain close to the background points. This indicates that the coupled correction does not substantially alter fields that are already close to the observations. This is a desirable behaviour, because it means that the machine-learning framework does not introduce unnecessary corrections in variables for which the atmospheric background is already near optimal.

For **surface soil moisture**, **precipitation**, **IASI T2m**, and **near-surface humidity**, the Taylor diagrams show a clearer displacement of the analysis points towards the observational reference. This indicates that the coupled analysis improves not only the overall error level, but also the representation of the observed variability. In other words, the PHASE1 correction moves the analysed fields closer to the observations both in amplitude and in phase. This effect is most evident in the **Bal** and **AO** configurations, which are the configurations that show the largest atmospheric gains.

Spatial ΔRMSE diagnostics

The spatial ΔRMSE maps provide the geographical distribution of the impact of the coupled analysis by showing the difference between the analysis RMSE and the reference RMSE at each observation location. Negative values therefore indicate that the machine-learning analysis improves on the reference, while positive values indicate a degradation.

For **surface soil moisture**, the ΔRMSE maps show widespread improvement over the European domain(not shown), consistent with the robust positive land skill obtained in all configurations. This confirms that the land benefit is not restricted to a few isolated locations, but is instead a large-scale feature of the proof-of-concept system.

For the **atmospheric variables**, the spatial signal is more heterogeneous. The largest and most coherent improvements are found for **precipitation**, **IASI T2m**, and **near-surface humidity** (not shown), in agreement with the skill-score results. By contrast, the maps for **u10m**, **v10m**, and **MSLP** show only weak changes, reflecting the limited space for improvement given the good ERA5 quality. Overall, the ΔRMSE diagnostics support the interpretation that the current GAT-Land-Atm framework is able to improve the variables that are most sensitive to land–atmosphere coupling, while leaving almost unchanged the fields whose background is already close to optimal.

Synthesis of the three representative configurations

To summarise the validation, three representative PHASE1 configurations were retained in order to describe the range of behaviour of the coupled system: a **land-optimal configuration**, a **balanced configuration**, and an **atmosphere-optimal configuration**. The land-optimal configuration maximises soil-moisture skill and gives the best land result, but with almost no atmospheric gain. The atmosphere-optimal configuration gives the strongest atmospheric improvement, especially for **precipitation**, **IASI T2m**, and **near-surface humidity**, but with a weaker land result. The balanced configuration provides the best compromise between the two components and is therefore the most relevant one for the project objective of building a coupled land–atmosphere analysis.

Overall, the validation shows that the machine-learning framework is able to improve the **land analysis in a robust way** and can also improve selected **atmospheric variables** that are strongly linked to land–atmosphere interactions. The clearest atmospheric gains are obtained for **precipitation**,, and **near-surface temperaure and humidity**, while little improvement is found for the **mean sea-level pressure**, whose initial error is already very small. These results demonstrate that the coupled methodology is scientifically promising and that the current proof-of-concept framework is able to transfer useful information between land and atmosphere. At the same time, the trade-off between land and atmospheric performance indicates that the present architecture is not yet fully optimised and that further work is needed to strengthen the coupled correction while preserving the land benefit.

Variable / diagnostic	SS LO	SS Bal	SS AO	Main result
Land soil moisture	**0.518**	**0.304**	**0.106**	Strong improvement in all configurations
u10m	0.000	0.000	0.000	No improvement
v10m	0.000	-0.001	0.000	No meaningful gain
MSLP	0.000	0.009		Very small positive impact only
T2m	0.001	0.028	0.026	Small positive impact
Precipitation	0.000	**0.121**	**0.134**	Positive impact

T2m IASI	0.001	**0.217**	**0.255**	Strong improvement
Near-surface humidity (IASI)	0.000	**0.208**	**0.214**	Strong improvement

Table 1: Skill Score for the 3 different training configurations

The computational cost and memory requirements for producing the GAT-AN analyses are summarised in Table 2.

		Period (years)	CPU Time	Max Memory GB
PHASE0	Train	4	27’	3.6
	Test	6	32”	2.1
PHASE1	Train	4	19’	3.3
	Test	2	25”	2.3
	Cycling	6	32”	

Table 2: CPU time and Memory used for 6 years of coupling DA

The current implementation requires only a single GPU with 80 GB of memory. By comparison, one global daily analysis cycle of SPREADS-Land, assimilating three observation types (surface soil moisture, leaf area index and snow depth) with 20 CPU nodes, takes approximately 15 minutes. This comparison highlights the potential computational efficiency of the machine-learning framework including training.

Conclusion

During this reporting period, the project successfully developed and demonstrated a first proof-of-concept machine-learning framework for coupled land–atmosphere data assimilation. The GAT-Land-Atm system shows that graph neural networks can combine land and atmospheric background information with observations to produce coupled analyses while requiring only a fraction of the computational cost of a traditional ensemble-based data assimilation system.

The validation demonstrates robust improvements for the land component and encouraging improvements for several atmospheric variables directly linked to land–atmosphere interactions, particularly precipitation, near-surface humidity and temperature-sensitive satellite observations. Although the current framework is still at the proof-of-concept stage, the results confirm that machine-learning-based coupled data assimilation is a promising approach and provides a solid basis for the next stages of the project.