



Novel approaches bridging EO and ESM data

Deliverable 1.1

Authors: Plésiat É., Beobide-Arsuaga G., Baehr J.,
Hurduc A., Ermida S., Dutra E., Trigo I., Verjans V.,
Donat M., Kuehnert J., Lessig C., Kadow C.

Document information

Grant agreement	101137656
Project title	EXPECT – Towards an integrated capability to explain and predict regional climate changes
Project acronym	EXPECT
Project start date	1 April 2024
Related work package	WP1 - Bridging Earth Observations and Earth System Models for new climate knowledge and improved predictions
Related task(s)	Task 1.1: Bridging different data types for data infilling, process evaluation and initialisation of predictions
Lead organisation	DKRZ
Authors	Plésiat É., Beobide-Arsuaga G., Baehr J., Hurduc A., Ermida S., Dutra E., Trigo I., Verjans V., Donat M., Kuehnert J., Lessig C., Kadow C.
Submission date	30/03/2026
Dissemination level	PU (Public)

History

Date	Submitted by	Reviewed by	Notes
27/02/2026	Plésiat Étienne		Initial version for internal review
11/03/2026	Kushner Paul		Internal review
12/03/2026	Stoffels Rikke		Internal review

Please cite this report as: Plésiat É., Beobide-Arsuaga G., Baehr J., Hurduc A., Ermida S., Dutra E., Trigo I., Verjans V., Donat M., Kuehnert J., Lessig C., Kadow C. (2026), Novel approaches bridging EO and ESM data, D1.1 of the EXPECT project

Disclaimer: 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 European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them.

Table of content

List of tables	5
List of figures	6
Abbreviations	7
Achievement table for DoA objectives	8
About	9
Executive summary	9
1. Introduction	11
2. Deep learning to fill data gaps in observation datasets	12
2.1 Introduction	12
2.2 Reconstruction of the HadEX-CAM Dataset	12
Dataset and motivation	12
Deep-learning based reconstruction method	13
Results and evaluation	13
2.3 Reconstruction of the GPCC Dataset	15
Dataset and motivation	15
Deep-learning based reconstruction method	16
Results and evaluation	17
3. Development and Application of Hybrid Machine-Learning Data Assimilation for Hindcast Experiments	21
3.1 Introduction	21
3.2 Hybrid Machine-Learning data assimilation	21
3.3 Hindcasts Initialized from Hybrid Machine-Learning Data Assimilation	23
4. EO based diagnostics of land-atmosphere coupling regimes	25
4.1 Introduction	25
4.2 Data	25
4.3 Methods	26
Regridding	26
Computing LST amplitude	26
Considering the incoming shortwave radiation	27
Computing hydrological regimes	28
4.4 Hydrological Regimes	29

5. Analogue-based decadal predictions	33
5.1 Introduction	33
5.2 Sensitivity analyses of analogue-based predictions	33
5.3 Including performance-based constraints in analogue selection	35
5.4 Multi-variate extension of the analogue method using Earth observations	38
6. ML extension of high-resolution reanalysis	42
6.1 Introduction	42
6.2 Methods	43
6.3 Results	43
7. Conclusions	47
7.1 Key results and impact for EXPECT	47
Contribution to the integrated EXPECT capability	48
Wider scientific and societal relevance	49
7.2 Contribution to next WPs and long-term project objectives	50
EXPECT Data availability	51
EXPECT Code repositories	51
EXPECT publications	51
References	53

List of tables

Table 1: Evaluation of the reconstruction methods using the test dataset (CMIP6 data).	13
Table 2: Same as Table 1 but evaluated using the ERA5 dataset remapped to the HadEX-CAM grid.	13
Table 3: Evaluation of the AI-based reconstruction methods using the test dataset (HighResMIP data) for seven evaluation metrics.	18
Table 4: Same as Table 3 for the ERA5 dataset.	18
Table 5: Performance scores comparing the model predictions with the baseline from the nearest neighbor interpolation calculated from a total of 240 samples from year 1999, i.e. the 20 first timesteps from each month.	43

List of figures

Figure 1: Linear trends (in days/10 years) of TX90p for the period 1901–2018.	17
Figure 2: Model schematic illustrating the encoder-processor-decoder architecture of the GNN model employed for the reconstruction of the GPCC dataset.	19
Figure 3: Illustration of the evaluation procedure for the GNN model using the ERA5 dataset.	20
Figure 4: a) Temporal mean RMSE of the GNN reconstruction of ERA5. b) Difference in temporal mean RMSE between the GNN and U-Net reconstructions.	22
Figure 5: Standard Kalman Filter assimilation and Hybrid machine learning (ML) assimilation of the North Atlantic Current's (NAC) northwest corner.	25
Figure 6: First lead year anomaly correlation coefficient (ACC) of global mean sea surface temperature anomalies relative to the 1985–2014 climatology, computed between ERA5 and hindcast initialized from standard Kalman Filter assimilation (Standard), and between ERA5 and hindcast initialized from hybrid machine-learning assimilation (ML).	26
Figure 7: First lead-year anomaly correlation coefficient (ACC) of Subpolar Gyre sea surface temperature anomalies relative to the 1985–2014 climatology, computed between ERA5 and hindcast initialized from standard Kalman Filter assimilation (Standard), and between ERA5 and hindcast initialized from hybrid machine-learning assimilation (ML).	27
Figure 8: Spatial distribution of dominant hydrological regimes over Europe.	32
Figure 9: Parameters associated with each regime.	33
Figure 10: Ratio (LSTamp/SWamp) versus soil moisture for the six pixels highlighted in Figure 8.	34
Figure 11: Fraction of global area with statistically significant added skill (5% confidence level) in near-surface temperature in the constrained compared to the unconstrained CMIP6 ensemble.	37
Figure 12: Sea-surface temperature deviance diagnostic.	39
Figure 13: Pseudo-observation hindcasting: skill scores of the weighted ensemble relative to the unweighted ensemble.	40
Figure 14: 1930–2024 hindcast performance. Evaluation of predicted SST average of hindcast years 1–10, verified against observations.	41
Figure 15: Hindcast prediction skill for SST (blue curves) and SSH (green curves) as a function of weights for initialization constraints varying individually between 0 and 1.	43

Figure 16: Hindcast prediction skill for SST (first two rows) and SSH (last three rows) for different configurations of the initialization constraints: constraining only on SST (left column), only on SSH (right column) and equally on SST and SSH (middle column)._ **44**

Figure 17: Architecture of the WeatherGenerator model._____ **46**

Figure18: Maps for 1000 hPa u-component wind for a single sample on 1999-10-01 6:00._____ **48**

Figure 19: Maps for 2m temperature for 1999-10-01 6:00._____ **49**

Abbreviations

ACC: Anomaly Correlation Coefficient
 ADW: Angular Distance Weighting
 AI: Artificial Intelligence
 AIC: Akaike Information Criterion
 CMIP: Coupled Model Intercomparison Project
 DJF: December, January, February
 EO: Earth Observation
 ESM: Earth System Model
 DL: Deep Learning
 JJA: June, July, August
 LSA SAF: Satellite Applications Facility on Land Surface Analysis
 LST: Land Surface Temperature
 MAE: Mean absolute error
 MAM: March, April, May
 ML: Machine Learning
 MSG: Meteosat Second Generation
 NAC: North Atlantic Current
 RMSE: Root mean square error
 RSS: Residual Sum of Squares
 SEVIRI: Spinning Enhanced Visible and Infrared Imager
 SM: Soil Moisture
 SON: September, October, November
 SPG: Subpolar Gyre
 SROCC: Spearman rank correlation coefficient
 SSTA: Sea surface temperature anomalies
 SW: Downward Surface Shortwave Flux

Achievement table for DoA objectives

DoA objective	Achievement in D1.1	Section in D1.1
Infilling incomplete observational datasets	Deep-learning reconstruction frameworks developed and applied to HadEX-CAM temperature extremes and GPCC precipitation datasets, producing spatially complete climate fields	Section 2: Deep learning to fill data gaps in observation datasets
Improved initialization of predictions	Hybrid machine learning data assimilation developed and applied to initialise MPI-ESM hindcasts	Section 3: Development and Application of Hybrid Machine-Learning Data Assimilation for Hindcast Experiments
Process-based evaluation using Earth Observation diagnostics	Earth-Observation-based analysis of land-surface temperature–soil-moisture coupling used to identify hydrological and land–atmosphere interaction regimes across Europe	Section 4: Temperature and soil moisture diagnostics of land-atmosphere coupling regimes for climate-model evaluation
Improved initialization of predictions	Analogue-based initialization with performance-based constraints implemented and evaluated	Section 5: Analogue-based decadal predictions
Data for extension of reanalysis datasets	Machine-learning framework developed to generate extended high-resolution regional reanalysis fields	Section 6: Machine Learning extension of high-resolution reanalysis

About

This deliverable (D1.1) is produced within Task 1.1 from Work Package 1 (WP1) of the EXPECT project. WP1 aims to develop methods that combine Earth Observation (EO) datasets and Earth System Model (ESM) information to improve understanding and prediction of regional climate variability and change.

Specifically, Task 1.1 develops methods to bridge EO and ESM data for infilling incomplete observational datasets, process-based evaluation of climate models using EO-derived diagnostics, improved initialization of climate predictions through new data assimilation and analogue approaches, and extension of reanalysis datasets. These activities address fundamental limitations in climate science arising from sparse historical observations, structural model biases, and inconsistencies between observational and model data.

D1.1 documents the methodological advances and applications developed in Task 1.1 across multiple components of the climate system. These include deep-learning-based reconstruction of temperature and precipitation observational datasets, hybrid machine-learning data assimilation for decadal hindcasts, EO-based characterization of land-atmosphere interactions, analogue-based prediction methods integrating EO constraints, and approaches to extend high-resolution reanalyses. Together, these developments demonstrate novel strategies to integrate EO and ESM information across spatial and temporal scales.

The results presented here provide new data products, methodological frameworks, and evaluation diagnostics that will support subsequent EXPECT work packages, particularly in attribution and prediction studies.

Executive summary

Understanding and predicting regional climate variability and climate change requires consistent integration of Earth Observations (EO) and Earth System Models (ESMs). Observational datasets are inherently incomplete and heterogeneous, while models are inherently limited in the processes they can represent and often exhibit biases. Addressing these challenges is a central objective of Work Package 1 (WP1) of the EXPECT project. Within WP1, Task 1.1 develops methods to bridge EO and ESM data by infilling incomplete observational datasets, enabling process-based

evaluation of climate models using EO-derived diagnostics, improving initialization of climate predictions through advancing data assimilation and analogue approaches, and enhancing reanalysis datasets.

This deliverable presents the methodological developments and applications achieved in Task 1.1. Deep-learning (DL) frameworks are used to infill incomplete observational records of temperature extremes and precipitation, yielding more accurate reconstruction of climate fields than traditional statistical interpolation methods. EO-based analyses of land-surface temperature and soil-moisture interactions provide diagnostics of land-atmosphere coupling regimes relevant for climate-model evaluation. Hybrid machine-learning (ML) data assimilation and analogue-based prediction methods demonstrate how observational constraints can improve initialization and predictive skill of decadal climate hindcasts. In addition, ML approaches are developed to extend high-resolution regional reanalysis information in space and time by learning relationships with coarser-resolution datasets.

Together, these results establish a coherent framework for bridging EO and ESM information across multiple components of the climate system that will support subsequent EXPECT work on climate attribution and prediction.

1. Introduction

Understanding and predicting regional climate variability and climate change requires consistent integration of Earth Observation (EO) and Earth System Models (ESM). Observations provide empirical information on the evolving state of the climate system, while models offer physically consistent representations of coupled processes and enable projections under changing external forcing. However, important structural differences between these data sources currently limit their combined use. Observational datasets are inherently incomplete and heterogeneous in space and time, particularly prior to the satellite era and in regions with poor station coverage. At the same time, ESMs exhibit systematic biases and imperfect representation of regional processes and variability.

Addressing these challenges is a central objective of the EXPECT project, which aims to develop an integrated capability to explain the factors and processes driving changes in climate and improve predictions on interannual to decadal timescales. Within this framework, Work Package 1 (WP1) focuses on bridging EO and ESM data to generate new climate knowledge and improved predictions. Task 1.1 specifically develops methods to bridge EO and ESM data by infilling incomplete observational datasets, enabling process-based evaluation of climate models using EO-derived diagnostics, improving initialization of climate predictions through data assimilation and analogue approaches, and extending reanalysis datasets:

- Deep Learning (DL) techniques enable reconstruction of incomplete observational climate datasets and are applied in this task to reconstruct records of temperature extremes and precipitation.
- EO-derived diagnostics from satellite and in-situ observations provide observationally constrained indicators of land-surface for model evaluation. In this task, analyses of land-surface temperature and soil-moisture observations are used to characterise land-atmosphere coupling regimes.
- To improve decadal hindcast initialization, hybrid Machine Learning (ML) data-assimilation and analogue-based frameworks are used to incorporate observational information into model states and ensemble selection.
- Finally, ML methods are also used to bridge datasets of differing spatial resolution, enabling extension of regional reanalysis information.

The datasets, diagnostics, and methodologies presented in this deliverable provide a foundation for subsequent EXPECT activities. In particular, they support WP4 activities on attribution and prediction through the developed analogue-based frameworks,

and WP6 activities on past and future hazard analysis using ML-infilled observations and complementary model-based datasets.

2. Deep learning to fill data gaps in observation datasets

2.1 Introduction

Reliable observational climate datasets are essential for understanding past variability, detecting long-term trends, and assessing future climate risks. However, many historical climate records suffer from substantial spatial and temporal gaps, particularly before the mid-20th century when observational networks were sparse. Traditional interpolation techniques such as inverse distance weighting (IDW) or kriging are commonly used to fill missing values, yet they often produce overly smooth fields. Recent advances in artificial intelligence (AI) have demonstrated strong potential for reconstructing missing geophysical data by learning spatial patterns from large datasets. In particular, neural network (NN)-based approaches have been used to reconstruct missing values in observational datasets by using *transfer learning*, i.e., by applying models trained on physically consistent climate model simulation data. Such approaches go beyond conventional statistical interpolation techniques in reconstructing climate fields such as temperature anomalies (Kadow et al., 2020). Supported by EXPECT, several DL approaches have been developed and tested at DKRZ to reconstruct two historical observational datasets: HadEX-CAM, which contains temperature-based extreme indices over Europe, and GPCC, which provides global precipitation observations.

2.2 Reconstruction of the HadEX-CAM Dataset

Dataset and motivation

The HadEX3 dataset provides land-based gridded indices of temperature and precipitation extremes derived from more than 30 000 meteorological stations worldwide at monthly and annual resolution from 1901-2018 (Dunn et al., 2020). These indices are calculated at stations and then interpolated onto a regular grid using angular distance weighting (ADW). While this interpolation produces spatially complete fields, it also introduces smoothing and potential biases, particularly in data-sparse regions and early decades. To enable AI-based reconstruction without prior interpolation, an intermediate product called HadEX-CAM was created for this

study. HadEX-CAM uses the same station-derived indices but applies the Climate Anomaly Method (CAM), which averages available station anomalies within each grid cell without spatial extrapolation. The resulting dataset preserves observed information but contains large gaps in space and time, making it suitable for testing reconstruction algorithms. It contains four temperature-extreme indices (TX90p, TN90p, TX10p, and TN10p) at monthly resolution across Europe, defined as follows:

- TN10p: % of days with daily minimum temperature below the 10th percentile
- TN90p: % of days with daily minimum temperature above the 90th percentile
- TX10p: % of days with daily maximum temperature below the 10th percentile
- TX90p: % of days with daily maximum temperature above the 90th percentile

Deep-learning based reconstruction method

To reconstruct HadEX-CAM, a convolutional neural network architecture based on a U-Net (Ronneberger et al., 2015) with partial convolutions was employed (Liu et al., 2018). Partial convolutions restrict each convolution to valid inputs and update the associated mask at every layer. This design enables effective inpainting of large and irregular regions of missing data, making the method well suited for climate datasets with heterogeneous observation coverage. As shown in Kadow et al. (2020), the U-Net architecture successfully demonstrated skill in reconstructing the HadCRUT4 global temperature anomaly dataset. Extending this approach, an updated reconstruction of HadCRUT5 for the most recent years has been achieved within the EXPECT project (Forster et al., 2025; Thorne et al., 2026). The model takes as input both the incomplete field and its missing-value mask and learns to predict values only where data is missing. Training relies on transfer learning from ESM simulations: 45 historical simulations from eight CMIP6 models (Eyring et al., 2016) were used to compute the complete extreme-index fields, to which HadEX-CAM masks were applied to create artificial missing values in the input data. The network made of 3 encoding layers and 3 decoding layers with 512 feature channels in the bottleneck layer was optimized using Adam optimizer with a mean absolute error (MAE) loss restricted to missing regions. Predictions were constrained to the extreme index range (0–100 %) using a Sigmoid function. Multiple models were trained and ensembled to reduce stochastic uncertainty.

Results and evaluation

The reconstructed HadEX-CAM dataset provides complete monthly fields of temperature-based extreme indices (TX90p, TN90p, TX10p, TN10p) over Europe from 1901–2018 with enhanced spatial detail relative to HadEX3. Evaluations against independent CMIP6 simulations and the ERA5 reanalysis (Hersbach et al., 2020) (Table 1 and 2) show that the AI reconstruction consistently outperforms statistical interpolation (IDW, kriging) and, in most cases, a generative approach based on the guided diffusion framework (Ho et al., 2020) across metrics such as the root mean square error (RMSE) and Spearman rank correlation coefficient (SROCC).

CMIP6	TX90p		TX10p		TN90p		TN10p	
Metrics	RMSE	SROCC	RMSE	SROCC	RMSE	SROCC	RMSE	SROCC
IDW	5.47	0.80	7.26	0.84	5.02	0.80	7.55	0.82
Kriging	5.12	0.81	6.96	0.85	4.77	0.81	7.12	0.84
U-Net	4.29	0.85	5.79	0.88	4.17	0.84	6.20	0.86
Diffusion	5.04	0.82	5.79	0.88	4.14	0.84	6.08	0.86

Table 1: Evaluation of the reconstruction methods using the test dataset (CMIP6 data). RMSE (in %, left inner column) and SROCC (right inner column) calculated on the reconstructed values only for each extreme index and for four datasets: the test datasets reconstructed using IDW, Kriging, U-Net, and diffusion models. RMSE is computed for each dataset and index across all spatial and temporal data combined, whereas SROCC is calculated for each time step individually and then averaged over time.

ERA5	TX90p		TX10p		TN90p		TN10p	
Metrics	RMSE	SROCC	RMSE	SROCC	RMSE	SROCC	RMSE	SROCC
IDW	5.52	0.82	5.83	0.84	5.26	0.83	6.3	0.83
Kriging	5.08	0.84	5.31	0.86	4.95	0.84	5.81	0.85
U-Net	4.39	0.87	4.7	0.88	4.33	0.87	5.24	0.87
Diffusion	4.71	0.86	4.88	0.88	4.36	0.87	5.39	0.86

Table 2: Same as Table 1 but evaluated using the ERA5 dataset remapped to the HadEX-CAM grid.

Further analysis reported in Plésiat et al., 2024 shows that the method is particularly effective in periods and regions with sparse observations, where traditional methods struggle. The reconstructed fields preserve large-scale means while revealing finer regional variability comparable to modern reanalyses.

For instance, as shown in Figure 1, the reconstruction of the TX90p index provides a much clearer picture of the long-term evolution of the frequency of warm days across Europe compared to the original HadEX3 dataset. While both datasets indicate an overall increase in warm-day frequency over the continent, the AI reconstruction reveals notable regional discrepancies, particularly in data-sparse areas such as the Mediterranean basin, as well as finer-scale spatial variability. These spatial patterns are in closer agreement with those typically observed in reanalysis products such as ERA5, although over a much longer historical period. This enhanced reconstruction provides a new basis for assessing climate-related risks and supports policy-relevant analyses at national and regional scales.

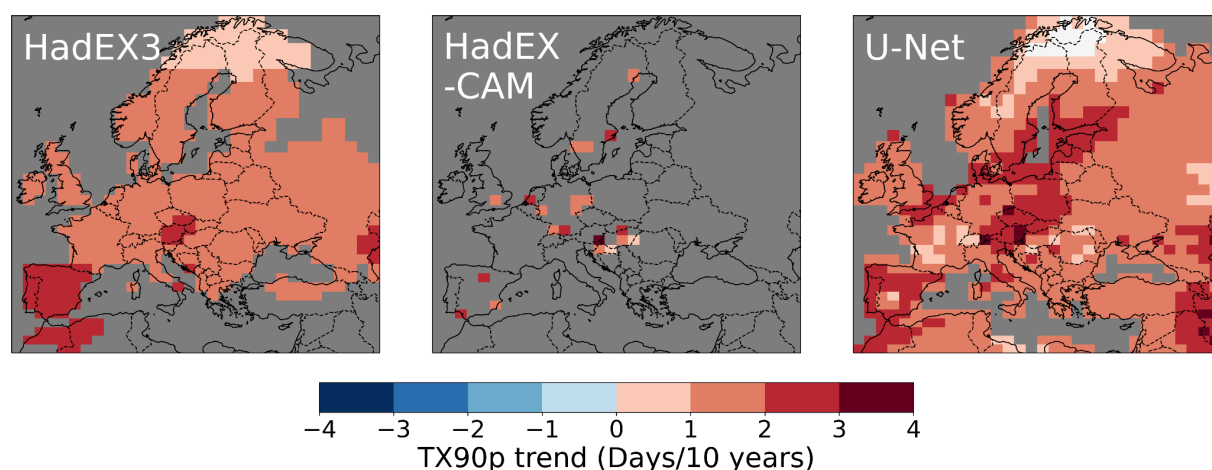


Figure 1: Linear trends (in days/10 years) of TX90p for the period 1901–2018. TX90p is the percentage of days when the daily maximum temperature is > 90th percentile. Left: original HadEX3 dataset, shown for grid boxes with at least 66% valid data over the full period. Center: original HadEX-CAM dataset, shown for grid boxes with complete temporal coverage. Right: U-Net reconstruction of the HadEX-CAM dataset.

2.3 Reconstruction of the GPCC Dataset

Dataset and motivation

The Global Precipitation Climatology Centre (GPCC) Full Data Monthly Version 2022 dataset provides monthly global land precipitation fields from 1891–2020 at different

spatial resolutions (Schneider et al., 2022). The reconstruction of the GPCC precipitation dataset poses substantially greater challenges than the HadEX-CAM case. First, GPCC is considered here at a higher spatial resolution (0.5°), implying a much larger number of grid cells and finer spatial variability that must be reconstructed consistently. Second, the proportion of missing values is also larger: while HadEX-CAM typically contains from 10 to 50 % valid data over land depending on the time period, the percentage of valid data over land does not exceed 27% for GPCC. Third, GPCC is a global dataset, in contrast to the regional European domain of HadEX-CAM. Conventional convolutional neural networks such as U-Nets exhibit several structural limitations for global climate applications. For instance, their convolution operators are formulated in Euclidean space and assume translation invariance, which does not strictly hold on the sphere. They further require fields defined on regular grid topologies, limiting their ability to operate directly on irregular spatial discretizations that may provide a more natural representation of global data. Moreover, the local nature of convolutions means that long-range spatial dependencies are captured only indirectly.

Deep-learning based reconstruction method

To address these limitations, we developed an encoder–processor–decoder Graph Neural Network (GNN) framework in which climate fields are represented on spherical meshes rather than planar grids (Lam et al., 2023).

In this approach, monthly precipitation fields defined on a regular latitude–longitude grid, together with positional embeddings, are encoded onto a hierarchy of multi-icosahedral meshes spanning multiple resolutions (M0–M5). Connectivity between grid and mesh nodes is defined through spatial cell overlap, such that each grid cell is linked to intersecting mesh cells. The edge features used in the encoder transformer-based graph convolution (Shi et al., 2020) comprise the great-circle distance, the relative direction and the normalized area overlap between the connected grid and mesh cells.

The processor propagates information across the multi-resolution mesh using 14 stacked local graph-transformer layers with a hidden feature dimension of 256. Message passing occurs on the mesh domain, linking nodes both across resolutions through hierarchical connections and within each mesh to their three nearest neighbours. In this stage, edge features are limited to the great-circle distance and relative direction between connected mesh nodes.

The decoder then projects the reconstructed fields back onto the target latitude–longitude grid using the same grid–mesh connectivity principles as in the encoder.

To improve convergence, the missing-value masks are appended as additional input features to provide the model with information about the missing-data pattern. Model parameters are optimized using the AdamW optimizer with a MAE loss weighted by normalized grid-cell area and applied only over masked regions. A schematic overview of the architecture is shown in Figure 2.

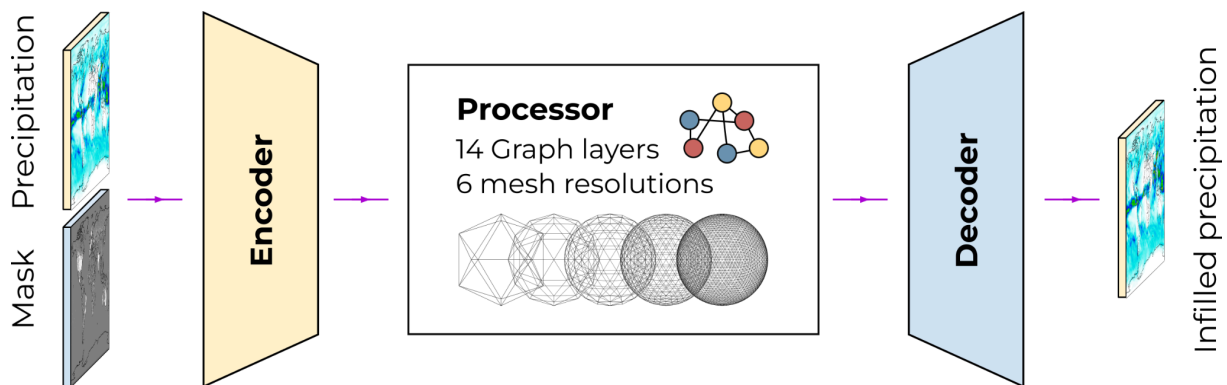


Figure 2: Model schematic illustrating the encoder-processor-decoder architecture of the GNN model employed for the reconstruction of the GPCC dataset.

By operating on a spherical mesh, this architecture more naturally represents geodesic relationships on the Earth's surface and enables more direct representation of long-range spatial dependencies than conventional grid-based convolutional networks.

Given the high spatial resolution and global domain, model training uses historical simulations from eight HighResMIP climate models to provide physically consistent complete precipitation fields. Realistic missing-value patterns are introduced by applying masks derived from the station-density information underlying the GPCC dataset, thereby generating training samples with observation gaps consistent with GPCC coverage.

Results and evaluation

Reconstruction skill is evaluated against both withheld HighResMIP simulations and ERA5 precipitation fields remapped to the GPCC grid. During evaluation, missing-value masks derived from the GPCC dataset are applied to the complete data at the corresponding dates to generate artificial gaps (Figure 3). The GNN reconstruction of the GPCC precipitation dataset is compared with the results obtained using the same U-Net framework as for the HadEX-CAM reconstruction but with a larger capacity (6 encoding/decoding layers and 2048 feature channels in the bottleneck layer). In this configuration, the U-Net was augmented with circular padding in longitude to enforce periodic boundary conditions and avoid artificial discontinuities at the dateline.

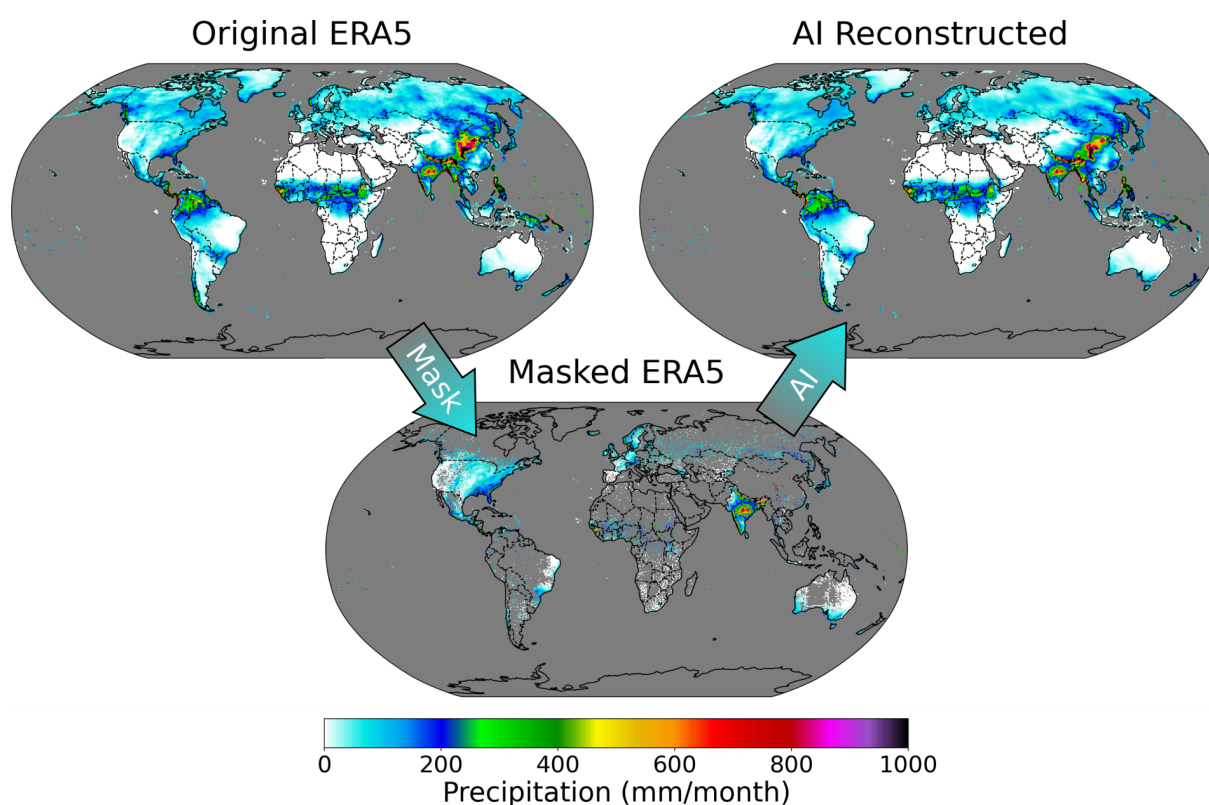


Figure 3: Illustration of the evaluation procedure for the GNN model using the ERA5 dataset. Top left: ERA5 monthly precipitation field for July 1946, remapped to the GPCC grid. Bottom center: masked version obtained by creating artificial missing values using the GPCC mask. Top right: AI reconstruction produced by the GNN model.

The reconstruction skill was evaluated using seven complementary metrics: root mean square error (RMSE), Spearman rank-order correlation coefficient (SROCC),

mean bias (Bias), coefficient of determination (R^2), spatial mean Wasserstein distance (SMWD), log-spectral distance (LSD), and the fractional skill score (FSS) for exceedances of the 90th percentile using a window size of 3x3 grid cells. Except for LSD, all metrics were computed as area-weighted quantities using cosine-latitude weights to account for the varying grid-cell area. As shown in Table 3 and 4, the two DL approaches yield very similar performance across all metrics and datasets, indicating that the graph-based model does not provide a clear accuracy advantage for this specific gridded GPCC reconstruction task. Results on ERA5 suggest marginally better generalization of the GNN, but the differences across metrics are too small to support conclusions about relative model performance.

HighResMip	RMSE	SROCC	Bias	R^2	SMWD	LSD	FSS
U-Net	29.97	0.97	-2.71	0.92	5.63	0.081	0.82
GNN	30.90	0.97	-2.52	0.91	5.25	0.064	0.80

Table 3: Evaluation of the AI-based reconstruction methods using the test dataset (HighResMIP data) for seven evaluation metrics.

ERA5	RMSE	SROCC	Bias	R^2	SMWD	LSD	FSS
U-Net	22.11	0.99	-4.12	0.95	7.11	0.021	0.86
GNN	21.30	0.99	-3.55	0.96	6.71	0.018	0.88

Table 4: Same as Table 3 for the ERA5 dataset.

It is also the case when looking at the temporally averaged RMSE difference map (Figure 4b) showing generally small-magnitude deviations that are spatially widespread. A slight predominance of negative values suggests marginally lower errors for the GNN in many regions, although the magnitude of these differences remains limited. An isolated area of higher positive values is visible near the Himalayan region, the origin of which is unclear. Overall, the spatial error characteristics of the two models remain very similar, with higher values in regions of high precipitation (e.g., the ITCZ belt) and in data-sparse areas with limited coverage (e.g., Greenland) (Figure 4a). Consistent with the metric-based evaluation, the spatial error distributions of the two models indicate comparable performance

across regions and no robust advantage of the GNN under low-coverage conditions for this particular single-variable.

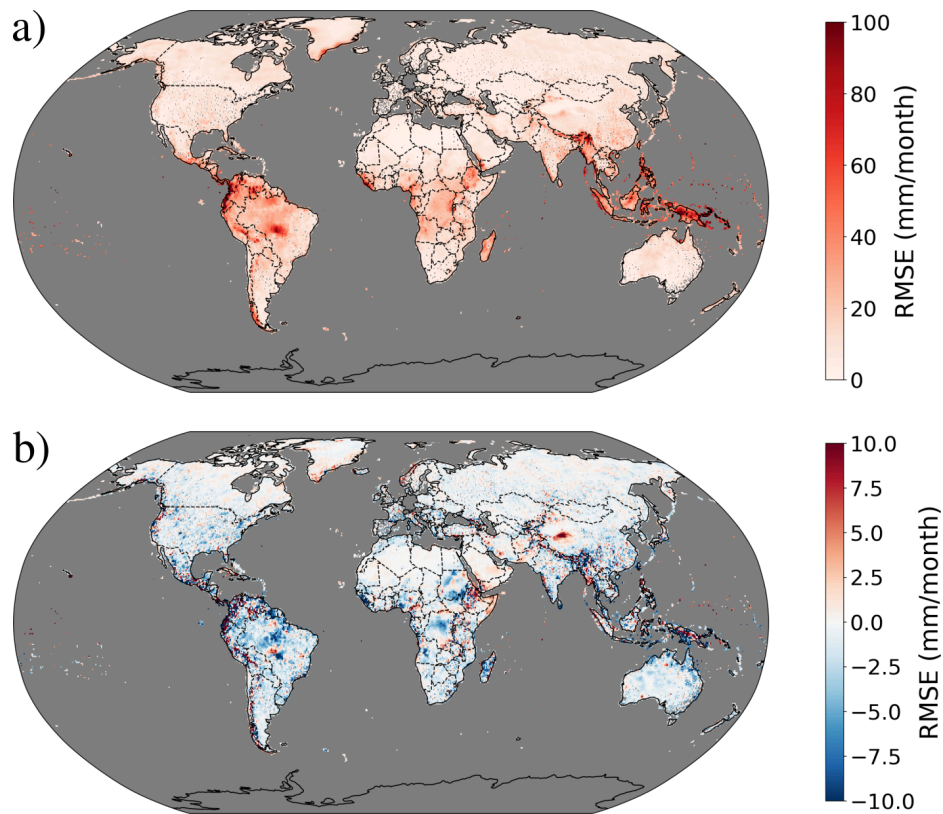


Figure 4: a) Temporal mean RMSE of the GNN reconstruction of ERA5. b) Difference in temporal mean RMSE between the GNN and U-Net reconstructions. Positive values indicate larger errors in the GNN reconstruction, while negative values indicate larger errors in the U-Net reconstruction. For both panels, only grid cells with at least 10 % reconstructed values are shown; remaining cells are masked in gray

Notably, however, the GNN is substantially more parameter-efficient (5M vs 63M trainable parameters), whereas the U-Net achieves significantly faster training (≈ 1 min vs 3 min per epoch) under comparable batch size, learning rate, and GPU configuration.

The reconstructed GPCC dataset is currently being prepared for public release to enable community analysis. Further work will investigate whether graph-based methods provide additional benefits when extending the framework to multivariate reconstruction or to data defined directly on spherical meshes or irregular station network.

3. Development and Application of Hybrid Machine-Learning Data Assimilation for Hindcast Experiments

3.1 Introduction

The University of Hamburg develops a hybrid machine-learning (ML) data assimilation approach to improve the initialization of climate hindcasts during periods of sparse observational coverage. The method combines state-of-the-art Ensemble Kalman Filtering (standard assimilation) with a neural-network-based reconstruction of North Atlantic Ocean heat content, trained during the well-observed Argo era (2004–2020) and applied to earlier decades (1960–2004). By transferring physically consistent ocean information from periods of dense observations to sparsely observed periods, this approach addresses an important limitation of standard assimilation, whereby poorly constrained EO initial conditions can lead to larger model biases and misrepresentation of critical ocean features. The methodology specifically targets biases in North Atlantic ocean structure that affect Subpolar Gyre (SPG) sea surface temperature (SST) variability and its coupling to the atmosphere, a dominant source of seasonal-to-interannual climate predictability for Europe and North America. This contribution documents the hybrid ML assimilation framework and evaluates its impact on initialized hindcast skill.

3.2 Hybrid Machine-Learning data assimilation

To overcome the limitations imposed by sparse historical observations, we developed a hybrid ML data assimilation framework (Lentz et al., 2025) that complements standard assimilation methods (Brune & Baehr, 2020). The approach consists of a neural-network-based reconstruction of North Atlantic Ocean heat content that is trained during periods with dense observational coverage (2004–2020, the Argo era) and applied to reconstruct ocean states during earlier, sparsely observed periods (1960–2004, pre-Argo era).

Specifically, we train a partial-convolution U-Net architecture as a supervised learning task using monthly ocean heat content fields from a traditional but state-of-the-art Ensemble Kalman Filter data assimilation. Training during the Argo era allows the network to learn physically consistent spatial patterns and gradients under

conditions of strong observational constraints. The trained model is then applied, via transfer learning, to reconstruct North Atlantic Ocean heat content for the pre-Argo period, targeting the upper 20 ocean layers that are most relevant for air–sea interaction and climate predictability.

The primary outcome of this hybrid ML assimilation is an improved representation of ocean heat content during sparsely observed periods. In particular, the method enhances the reconstruction of the northwestern region of the North Atlantic Current, a dynamically important area that is known to be challenging for coupled Earth system models and is closely linked to SST biases in SPG (Lentz et al., 2025). This improvement is illustrated in Figure 5, where the hybrid ML reconstruction exhibits more realistic spatial structure and gradients compared to standard assimilation.

These improved ocean states are especially relevant for climate predictability, as SPG SST variability plays a central role in modulating ocean–atmosphere coupling in the North Atlantic.

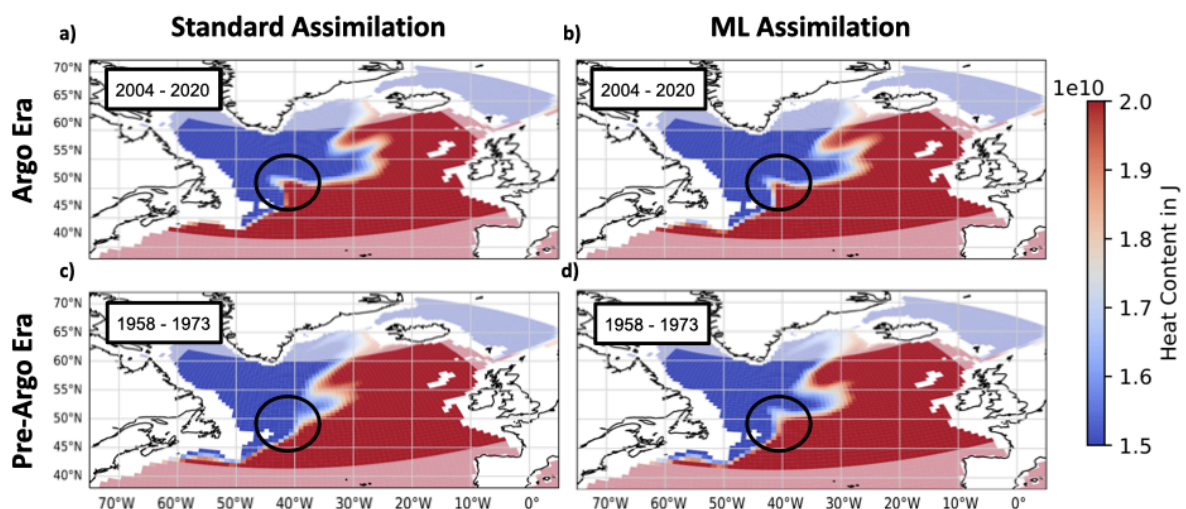


Figure 5: Standard Kalman Filter assimilation and Hybrid machine learning (ML) assimilation of the North Atlantic Current's (NAC) northwest corner. All ocean heat content estimates are calibrated to show the NAC flow as the gradient between the red and blue areas, while the black circles show the region of the NAC northwest corner. During the Argo era (2004–2020): a) Standard assimilation; b) ML assimilation. Pre-Argo era (1958–1973): c) Standard assimilation; d) ML assimilation.

3.3 Hindcasts Initialized from Hybrid Machine-Learning Data Assimilation

To assess whether the improved initial ocean states translate into enhanced climate prediction skill, we performed a comprehensive set of initialized hindcast experiments using the MPI Earth System Model decadal prediction system. MPI-ESM represents the atmosphere, ocean, sea ice, and land surface, with atmospheric processes simulated by ECHAM, ocean dynamics by MPIOM, and land-surface and biosphere processes by JSBACH. The system is run at low resolution, with T63 horizontal resolution and 47 vertical levels in the atmosphere, approximately 1.5° horizontal resolution with 40 vertical levels in the ocean, and monthly temporal resolution.

We conducted 16-member ensemble hindcasts of 26-month duration, initialized each November from 1960 to 2022. For the hybrid ML experiments, the initial conditions are identical to those used in the state-of-the-art Ensemble Kalman Filter initialization, except that the upper 20 ocean layers of North Atlantic Ocean heat content are replaced by the ML-reconstructed fields described in section 3.2. This design ensures that any differences in prediction skill can be directly attributed to the modified ocean initial conditions.

Prediction skill is evaluated using the anomaly correlation coefficient (ACC) relative to ERA5, with anomalies defined with respect to the 1985–2014 climatological reference period. Our results show that, for global and annual mean sea surface temperature anomalies (SSTAs), the overall prediction skill for the first lead year is comparable between the hybrid ML-initialized and standard initialized hindcasts in both the sparsely observed period (1965–1995; Figure 6a) and the well-observed period (1990–2020; Figure 6b). The comparable hindcast's skill indicates that the hybrid ML approach does not introduce spurious degradation elsewhere in the system.

In contrast, we find a clear improvement in SPG mean SSTA prediction skill during the first lead year, particularly in winter and spring, for the sparsely observed period (1965–1995; Figure 7a). In the first winter after initialization, ACC values for SPG SSTA increase from below 0.6 in the standard initialized hindcasts to above 0.8 in the hybrid ML-initialized simulations. Similar improvements are found in the first predicted spring. No significant changes are observed during the later, well-observed period (1990–2020; Figure 7b), as expected, since standard and hybrid ML assimilations converge when observational constraints are already strong.

These results indicate that hindcasts initialized during periods of sparse observational coverage may underestimate the role of ocean initial conditions in shaping near-term climate evolution. The enhanced SPG SSTA prediction skill obtained here suggests the potential for improved European temperature prediction skill, given the well-established link between SPG SSTAs, their coupling to the North Atlantic Oscillation, and downstream impacts on European surface air temperatures. The impact of improved SPG SSTAs on European surface air temperature prediction skill, together with the physical mechanisms linking ocean initialization to atmospheric predictability, will be evaluated in Task 2.3, ‘*Novel data assimilation for climate prediction*’.

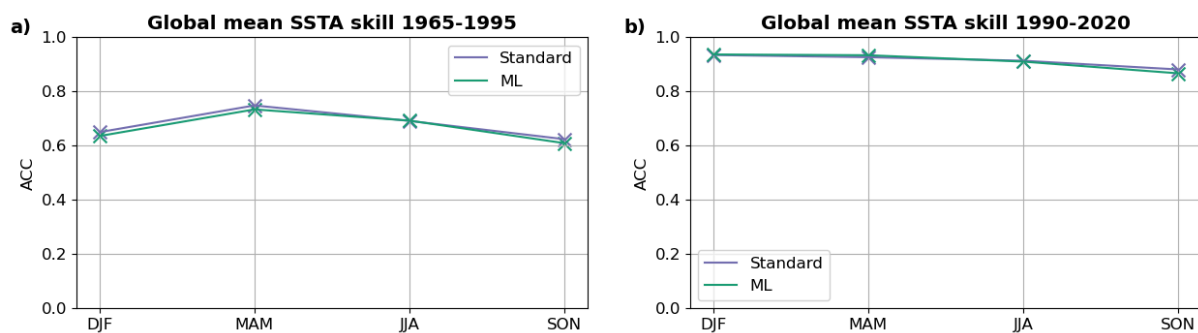


Figure 6: First lead year anomaly correlation coefficient (ACC) of global mean sea surface temperature anomalies relative to the 1985–2014 climatology, computed between ERA5 and hindcast initialized from standard Kalman Filter assimilation (Standard), and between ERA5 and hindcast initialized from hybrid machine-learning assimilation (ML). Sea surface temperatures are averaged globally and by season—winter (DJF), spring (MAM), summer (JJA), and autumn (SON)—prior to ACC calculation. Results are shown for a) 1965–1995; b) 1990–2020 period. Crosses indicate statistically significant skill at the 95% confidence level.

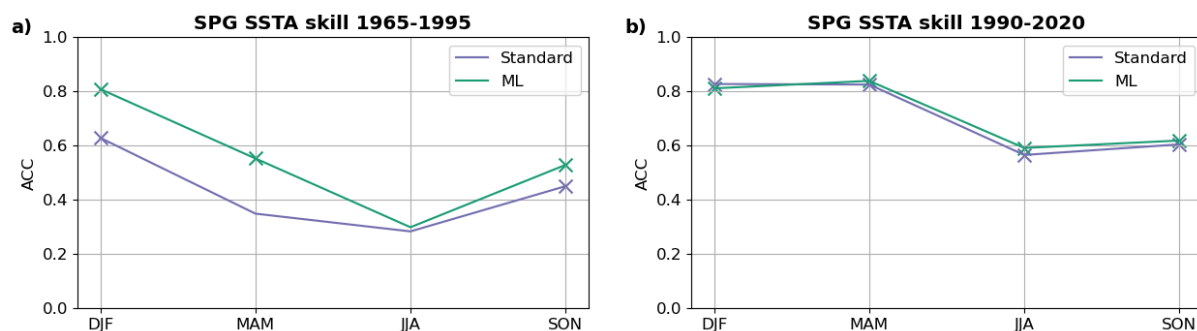


Figure 7: First lead-year anomaly correlation coefficient (ACC) of Subpolar Gyre sea surface temperature anomalies relative to the 1985–2014 climatology, computed between ERA5 and hindcast initialized from standard Kalman Filter assimilation (Standard), and between ERA5 and hindcast initialized from hybrid machine-learning assimilation (ML). Sea surface temperatures are averaged over the Subpolar Gyre region (50°–66°N, 60°–10°W) and by season—winter (DJF), spring (MAM), summer (JJA), and autumn (SON)—prior to ACC calculation. Results are shown for a) 1965–1995; b) 1990–2020 period. Crosses indicate statistically significant skill at the 95% confidence level.

4. EO based diagnostics of land-atmosphere coupling regimes

4.1 Introduction

The IPMA team studied the relationship between land surface temperature (LST) and soil moisture (SM), to identify hydrological regimes based on EO. Data at the pixel scale over the European continent covered the period from 2004 to 2024. This analysis provides a proxy for land-atmosphere feedback and can be further used in improving land-surface model behavior during periods of increasing evaporative stress.

4.2 Data

The LST product is developed by Satellite Applications Facility on Land Surface Analysis (LSA SAF) based on EO from the Spinning Enhanced Visible and Infrared Imager (SEVIRI) sensor onboard Meteosat Second Generation 0° satellite (MSG). The dataset is available every 15 minutes under clear sky conditions (Trigo et al., 2011), with 3 km resolution at nadir.

The SM product corresponds to the daily volumetric soil moisture from the Copernicus Climate Change Service (C3S CDS Soil Moisture 2018, v202312.0.0) at a resolution of 0.25° x 0.25°.

As an auxiliary variable the downward surface shortwave flux (SW; Trigo et al., 2011) from the SEVIRI/MSG was also used, with the same spatio-temporal characteristics of the LST dataset.

All datasets cover the period from 2004 to 2024.

4.3 Methods

Regridding

As the datasets are based on different EO sensors, their spatial resolution is also different. A conservative regridding methodology was employed to regrid the SEVIRI based datasets to the coarser SM resolution. This method preserves the total quantity when transferring between fields, as it computes fractional overlaps between cells and uses them to perform an area weighted distribution. The methodology goes as follows:

- The coarse pixel receives the weighted average of all overlapping SEVIRI pixels.
- The weight of each fine pixel is the fraction of its area inside the coarse pixel.
- The sum of weighted contributions is the area-average for that coarse pixel.
- **Threshold:** only consider coarse pixels that are based on at least 50% of the total possible fine pixels.

Computing LST amplitude

One way to assess the impact of SM on LST is through the diurnal amplitude of the latter (LST_{amp}). This amplitude is calculated as the difference between the daily maximum and minimum LST for each pixel.

The identification of the time of day when these temperatures occur depends on several factors such as local solar time (dependent on location), type of land cover, etc. To obtain the most likely time of occurrence of the minimum and maximum, the product of monthly averages, produced within the LSA SAF, was used. For each pixel belonging to the European continent and for each month throughout the period 2004 – 2024, the time of day when the surface temperature was minimum/maximum was identified. Then, for each month, the average was calculated for the period 2004 – 2024, resulting in monthly values. Monthly times of minimum and maximum temperature were used in order to decrease the computational cost of computing them based on the 96 daily values along the 21 year period.

The LST_{amp} was calculated at the daily level as:

$$LST_{amp} = LST_{max} - LST_{min} \quad (1)$$

Where,

$$LST_{max} = mean(LST_{t-1}, LST_t, LST_{t+1}) \quad (2)$$

$$LST_{min} = mean(LST_{t-2}, LST_{t-1}, LST_t, LST_{t+1}, LST_{t+2}) \quad (3)$$

Both LST_{max} and LST_{min} were only considered if they were based on at least three valid observations. More timesteps were considered for the computation of LST_{min} to account for the larger cloud contamination at this time of day.

Considering the incoming shortwave radiation

Values of LST_{amp} are highly dependent on the time of year, as such, LST_{amp} is normalized by the SW_{amp} . The SW_{amp} corresponds to the value of SW_{max} at the same time of occurrence of LST_{max} , since the SW_{min} is always equal to zero. The rationale behind this normalization goes as follows. If the LST_{amp} is the difference between the LST at peak radiation time and nighttime, assuming that there is no input SW radiation and the latent heat flux (LE) is approximately null at nighttime:

$$\text{Night: } \varepsilon \sigma T_0^4 = \varepsilon LW_0 - H_0 - G_0 \quad (4)$$

$$\text{Day: } \varepsilon \sigma T_{mx}^4 = (1 - \alpha)SW_{mx} + \varepsilon LW_{mx} - LE_{mx} - H_{mx} - G_{mx} \quad (5)$$

where, ε is the surface emissivity, σ is the Stefan–Boltzmann constant, T is the temperature, LW is the downwelling surface longwave radiation, H is the sensible heat flux, G is the ground heat flux, α is the surface albedo and LE is the latent heat flux. The subscripts $_0$ and $_{mx}$ are relative to the nighttime and daytime period, respectively.

Subtracting day from night:

$$\varepsilon \sigma (T_{mx}^4 - T_0^4) = (1 - \alpha)SW_{mx} + \varepsilon \Delta LW - LE_{mx} - \Delta H - \Delta G \quad (6)$$

After linearizing the first term:

$$\varepsilon \sigma T_{med}^3 \Delta T = (1 - \alpha)SW_{mx} + \varepsilon \Delta LW - LE_{mx} - \Delta H - \Delta G \quad (7)$$

Resulting in:

$$\frac{\varepsilon \sigma T_{med}^3 \Delta T}{(1 - \alpha)SW_{mx}} = \frac{1 + \varepsilon \Delta LW}{(1 - \alpha)SW_{mx}} - \frac{LE_{mx} + \Delta H + \Delta G}{(1 - \alpha)SW_{mx}} \quad (8)$$

$$\Delta T = \frac{(1 - \alpha)SW_{mx}}{\varepsilon \sigma T_{med}^3} \quad (9)$$

In this study we consider that as SW dominates the daily variability, the remaining terms are second order corrections that can be simplified at this stage. Normalizing LST_{amp} by shortwave radiation provides a first-order diagnostic of surface energy partitioning, with deviations reflecting changes in evaporative cooling efficiency.

Computing hydrological regimes

The transition between energy limited and water limited regimes is identified based on dry-down periods (periods where no additional moisture was added to the soil), as defined in Feldman et al. (2019). This is being done to ensure a larger SM range encompassing both energy limited and water limited regimes (when they exist). Dry-down periods are identified as follows:

- Days with decreasing SM relative to the previous day.
- The decrease must persist for at least 5 consecutive days (to ensure surface drying) and no more than 20 consecutive days (to avoid incorporating seasonal behaviors).
- The day preceding the start of each dry-down must have a SM value at least $0.01 \text{ m}^3 \text{ m}^{-3}$ higher, as an indication of rain.

Finally, the LST_{amp}/SW_{amp} ratio (hereafter mentioned as ratio, for simplification) is then analyzed together with the daily SM values to determine pixelwise changes in hydrological regimes.

To robustly estimate the relationship between SM and the ratio, and thereby identify the hydrological regimes, we used a binning and resampling methodology. To this end, SM observations were first grouped into bins of width $0.02 \text{ m}^3 \text{ m}^{-3}$. From each bin, 10 observations were randomly selected and combined across all bins to construct a

sample used to estimate the relationship between SM and the ratio. Bins with less than 10 total observations were not considered. This procedure was repeated 60 times. Three hydrological regimes were fitted at the pixel-level:

- an energy limited regime, where the LST_{amp}/SW_{amp} ratio remains constant with increasing SM.
- a water limited regime, where the LST_{amp}/SW_{amp} ratio decreases with increasing SM.
- transitional one, where a breakpoint between the two previous regimes can be found.

All three regimes were computed pixelwise based on the mean of all 60 bootstraps.

To identify the best fitting model, the same binning and bootstrapping methodology was used to compute the residual sum of squares (RSS), used for the Akaike Information Criterion (AIC), to minimize the error and model complexity.

$$AIC = 2k + N\ln(RSS) , \quad (10)$$

where k is the number of parameters for each regime/model and N is the number of samples (i.e., pairs of ratio and SM).

4.4 Hydrological Regimes

The resulting regimes are shown in Figure 8. The energy limited regimes are more prevalent in the northern part of the European continent along with the mountainous regions (Alps, Apennines, Balkans, Carpathians). The water limited regimes are more frequent in the southern part of the continent and the valley topography such as over northwestern France, Poland and Hungary. The transitional regime tends to concentrate over northern Africa and the Syrian Desert.

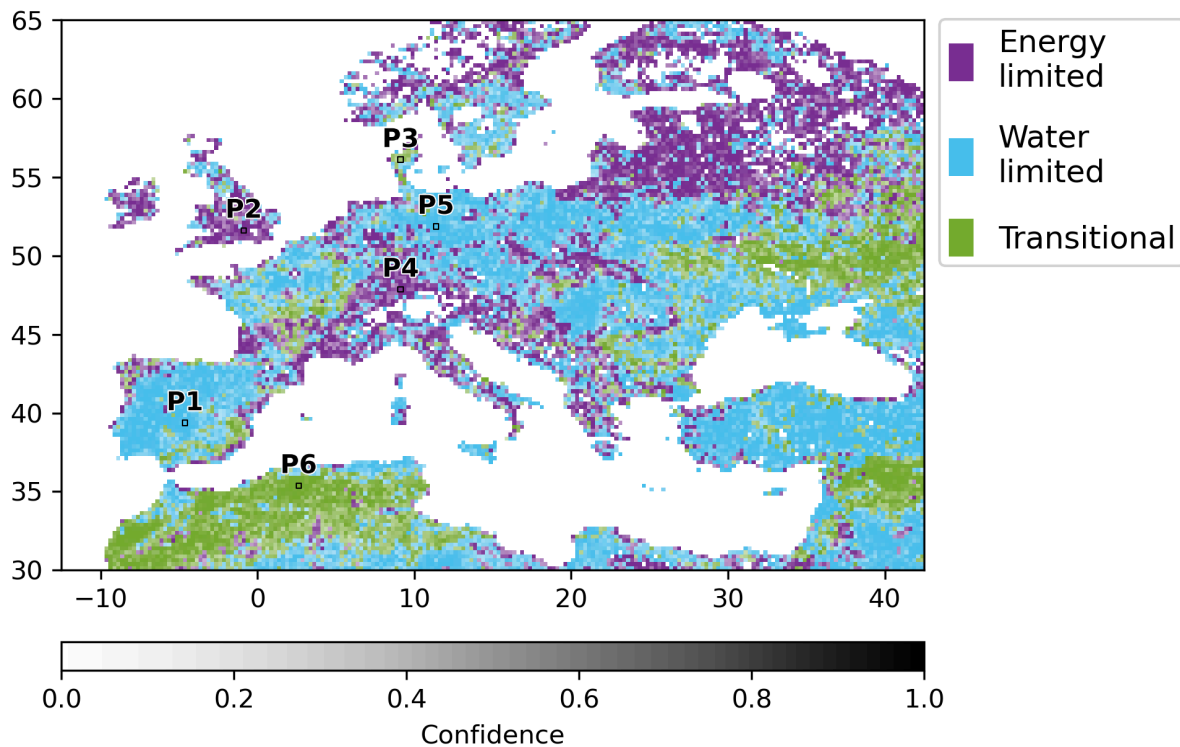


Figure 8. Spatial distribution of dominant hydrological regimes over Europe. Color opacity relates to the confidence in the model. Highlighted pixels are further analysed in Figure 10.

The parameters associated with each regime are shown in Figure 9. Regarding the energy limited regime, the intercept value (Figure 9a) corresponds to the mean ratio value. Lower values are observed over northern Europe and mountain regions, with a strong latitudinal gradient. As expected, southern Europe and northern Africa show higher values as the amount of available radiation and consequently temperature is higher.

Regarding the water limited regime, intercept values (Figure 9b) remain smaller in the northern part of the continent and mountain regions. Slope values (Figure 9c) are mainly negative, indicating a decrease in temperature with increasing SM.

For the transitional regime, regarding the slope preceding the breaking point (Figure 9e), values are generally small (-0.1 to 0.1) indicating energy limitation. After the breakpoint these regions show a strong negative slope, denoting water limitation. The spatial distribution of the breaking point (Figure 9g) shows a clear south to north gradient, implying that the transition between regimes occurs at lower SM values in the south. This can be due to the sparse vegetation in the south, where LST amplitude becomes water limited quickly and at low values of SM.

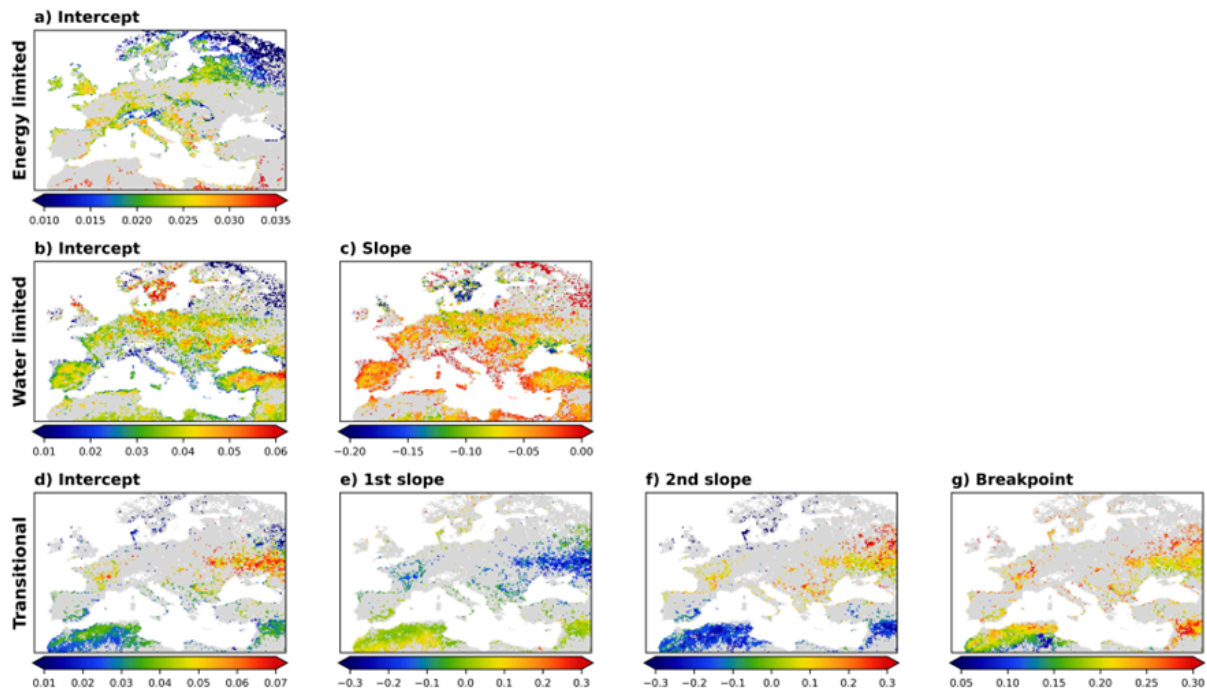


Figure 9. Parameters associated with each regime.

As an example, six pixels were selected (Figure 10; highlighted in Figure 8). Note that the x-axis represents increasing soil moisture, while the time evolution is in the decreasing soil moisture direction. During the energy limited regime (Figures 10b, d), ratio values are somewhat dispersed with no apparent relationship with SM values. For the water limited regime (Figures 10a, e), the negative slope is expected as increase in SM is associated with a decrease of LST amplitude. For the transitional regime (Figures 10c, f), preceding the breakpoint and at lower SM values the slope is positive/weak. After the breakpoint, the slope becomes negative and steeper, coinciding with water limited conditions.

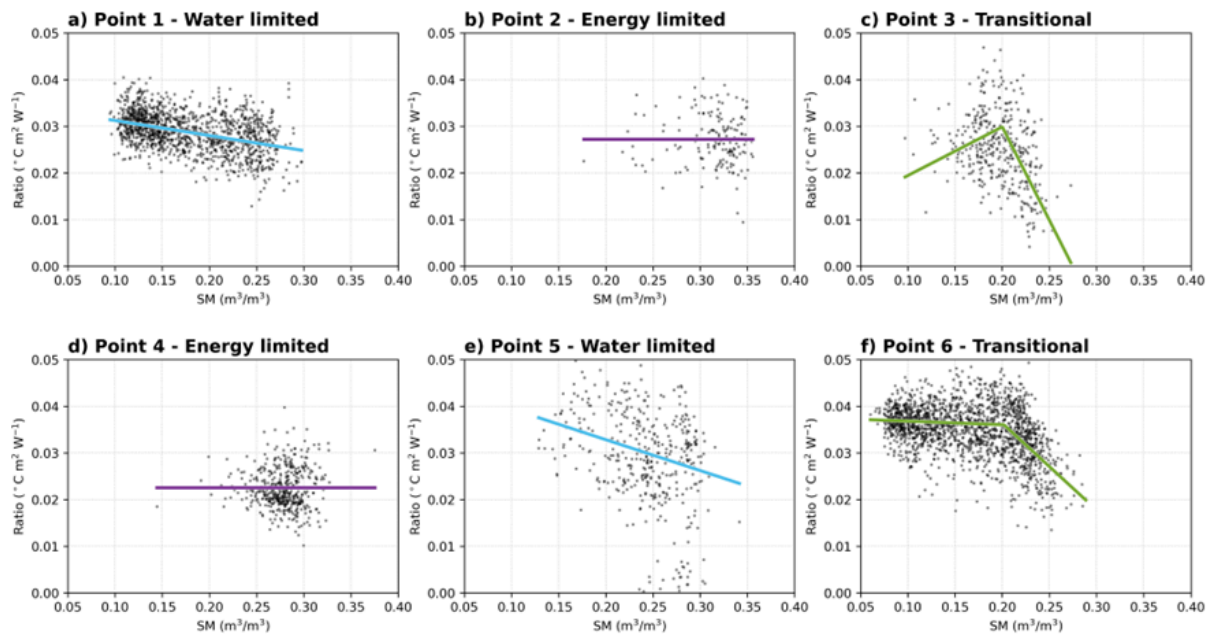


Figure 10. Ratio (LST_{amp}/SW_{amp}) versus soil moisture for the six pixels highlighted in Figure 8.

The complex interactions between LST and SM are hard to pinpoint. In this study, EO datasets were used to characterize the relationship between these two variables along a 21-year period, at a pixel level. The analysis carried out here is fully based on EO datasets, without the additional need of considering model simulations, reflecting the physical relationship between the variables studied.

As different regions follow different hydrological regimes, information about the dominant local regimes may prove useful in improving model simulations, as a proxy for land-atmosphere feedback. The identified drydown-based LST–SM regimes provide physically interpretable benchmarks for evaluating land-surface model behavior during periods of increasing evaporative stress. In particular, regime transitions and slope changes constrain the timing and strength of modeled evapotranspiration limitations, which are known sources of uncertainty in current models.

Here, the LST–SM relationship was identified based on the overall 2004 – 2024 period. This relationship may change during extreme temperature and precipitation events. The presented methodology will be used for work to be developed in task 5.2, where temperature and precipitation extremes are going to be analyzed based on SM and vegetation interactions.

Ultimately, several other factors affect the LST-SM relationship, such as vegetation phenology, type of soil, albedo, atmospheric conditions, etc. The influence of these factors is outside the scope of the present study.

5. Analogue-based decadal predictions

5.1 Introduction

The BSC has developed an analogue-based method to constrain decadal-scale internal climate variability in climate model projections. This method selects simulations from a large ensemble of climate projections that are more similar to the observed state at the time of starting the predictions (Mahmood et al. 2022). This approach has been demonstrated to improve the skill for near-term (multi-annual, decadal, and multi-decadal) climate predictions by capturing both the response to external forcing and the phase of internal climate variability.

The analogue-based prediction method has several advantages compared to more sophisticated initialized decadal prediction systems that use data assimilation methods. First, analogue-based predictions require only a fraction of the computational expense of initialized predictions, since the latter need frequent re-initializations of model ensembles from observations. Second, initialized predictions can be subject to spurious adjustments after initialization, as the model transitions from the observed state to a model-specific dynamical state space, known as model drift. However, analogue-based predictions also have their limitations. They require a large set of existing simulations to find analogues matching observations closely at the forecast start time. Furthermore, they rely on the assumption that the selected analogues are representative of the entire climate system, while any selection method is necessarily based on low-dimensional diagnostics.

5.2 Sensitivity analyses of analogue-based predictions

We have further explored and developed this analogue-based prediction method by studying the sensitivities to the metric used for analogue selection. In particular, we compared two selection metrics: the uncentered and centered pattern correlations. These pattern correlations are both computed between model simulations and observations at the forecast start date to select the best-matching simulations as analogues. The uncentered pattern correlation considers the patterns of the global SST anomalies including the signatures of a warming trend over the recent decades. In contrast, the centered pattern correlation removes the effect of

the warming trend by subtracting the global mean from regional anomalies every year. We found that constraining based on the uncentered pattern correlation significantly improves predictive skill in 1960-2020 hindcasts over substantially larger areas compared to the centered pattern correlation constraint (Figure 11).

However, the results are more nuanced when examining regional skill improvements, as both constraining methods show skill gains in different regions. The uncentered constraint leads to overall similar skill patterns as decadal predictions with more complex initialized prediction systems. In contrast, the centered constraint improves the hindcast quality primarily in the North Atlantic and parts of the North Pacific Ocean - two regions characterized by pronounced decadal climate variability. These findings highlight that the centered constraint focuses on predicting variations associated with internal climate variability, whereas the uncentered constraint also incorporates a strong effect from the global warming trend. This research has been published, and more details can be found in Donat et al. (2024).

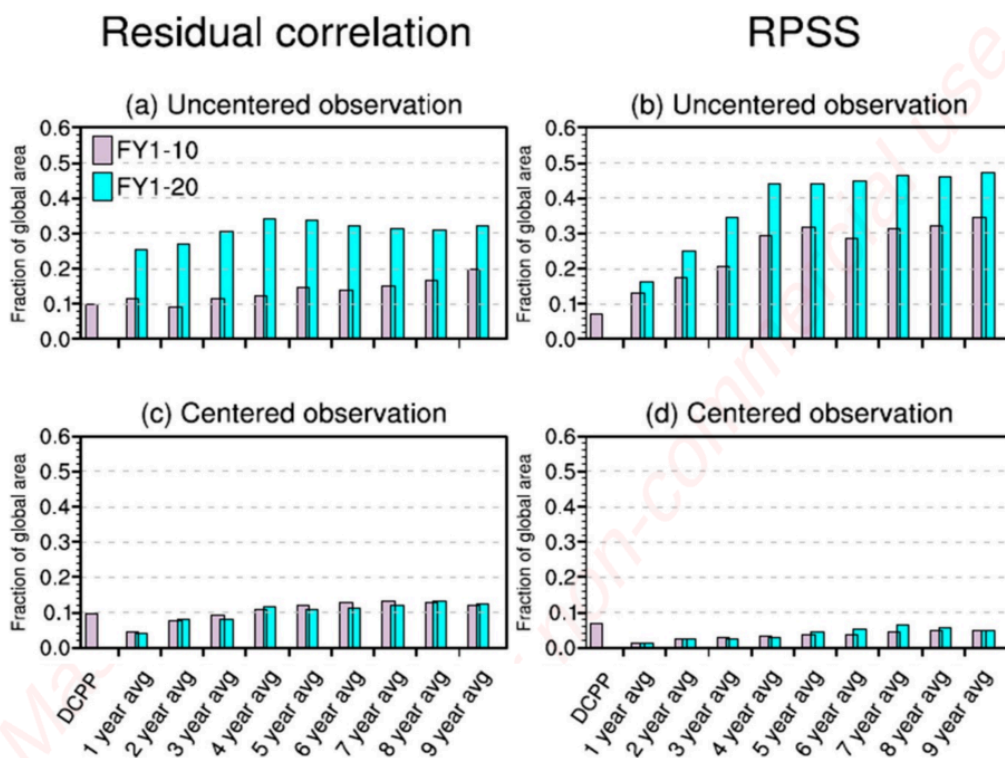


Figure 11: Fraction of global area with statistically significant added skill (5% confidence level) in near-surface temperature in the constrained compared to the unconstrained CMIP6 ensemble. The skill is measured as residual correlation in a, c, and as ranked probability skill score (RPSS) in b, d. In each panel the added skill is shown for constraints using SST anomaly patterns averaged over 1–9 years prior to hindcast start year. For reference, the area with added skill in an ensemble of initialized decadal

hindcasts (the Decadal Climate Prediction Project, DCP) is shown on the left in each subpanel. Light pink shows evaluation of hindcast years 1–10 average, and blue of hindcast years 1–20 average. The evaluations are based on the 40 constraints applied in all hindcasts with start years ranging from 1961 to 2000. Figure adapted from Donat et al. (2024).

5.3 Including performance-based constraints in analogue selection

Analogue-based initialization selects initial states matching observations from large climate model ensemble simulations, but neglects criteria of model performance. In further work, we have implemented performance-based model weighting, favoring analogues from models that show better consistency with observations in climate forced response and stationary dynamics. This work first required establishing an objective metric to quantify consistency with observations. Focusing on SST decadal predictions, we have implemented a systematic framework to compute a deviance statistic, not previously used in model weighting schemes. The deviance is a statistical metric, capturing model performance in different components of SST spatiotemporal evolution: response to radiative forcing, predictable dynamics, and noise covariance structure. We have therefore investigated its use as an objective criterion in model weighting to improve analogue-based predictions.

We have first computed the SST deviance for 39 climate models, using a total of 523 simulations. Results are shown in Figure 12 for the total deviance metric (a), and the sub-deviance components (b, c, d). Higher values indicate stronger model-observation inconsistencies, with the value of 1 corresponding to statistically significant deviance at the 1% confidence level. All models show significant total deviance with respect to observations, dominated by the noise covariance sub-deviance, followed by the predictable dynamics sub-deviance, and by the radiative forcing response sub-deviance. We find inter-model variability in every sub-deviance component, indicating that model differences result in statistically measurable differences in agreement with observed SST spatiotemporal evolution.

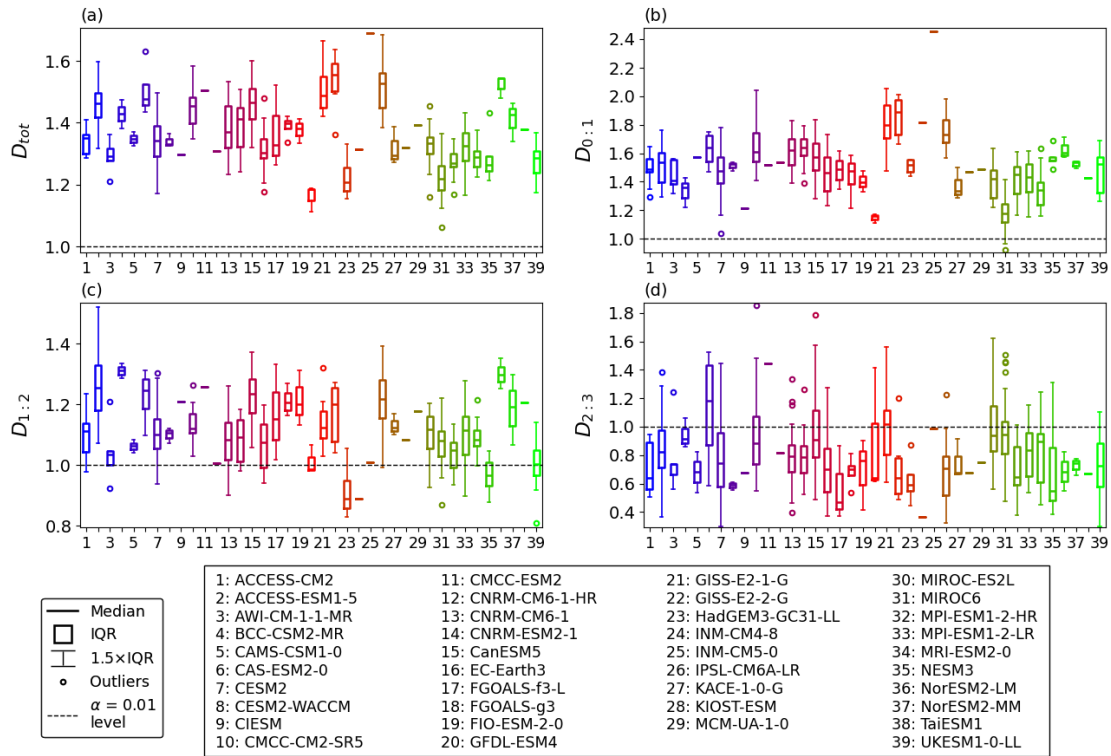


Figure 12: SST deviance diagnostic. (a) Total deviance (D_{tot}), (b) noise covariance deviance ($D_{0:1}$), (c) predictable dynamics deviance ($D_{1:2}$), and (d) radiative forcing response deviance ($D_{2:3}$) statistics from 39 CMIP6 models with respect to observations (1930-2024). Boxplots show spread from different ensemble members. Higher D indicates stronger statistical inconsistency with observations, with the value 1 corresponding to statistical significance at the 1% level (dashed horizontal lines). Climate model names corresponding to numbers in (a,b,c,d) are given in the legend below. IQR: inter-quartile range. Figure from Verjans et al. (2026).

To calibrate the deviance as a weighting metric for analogue selection, we use a pseudo-observation prediction framework. In this framework, we target model realizations to be predicted using all other climate models from the ensemble. This enables a robust calibration, since the prediction exercise can be repeated with each simulation as a target pseudo-observation, and therefore reducing sampling uncertainties in the calibration. We have performed a thorough search of parameter space to optimize the weights to be applied to each sub-deviance component in the analogue selection, as well as the balance between these performance criteria and the initialization constraint to match internal climate variability at hindcast start time. Our results show that, when the deviance weight is appropriately calibrated for analogue selection, the performance-based weighted ensemble performs more skillful predictions than the unweighted ensemble (Figure 13).

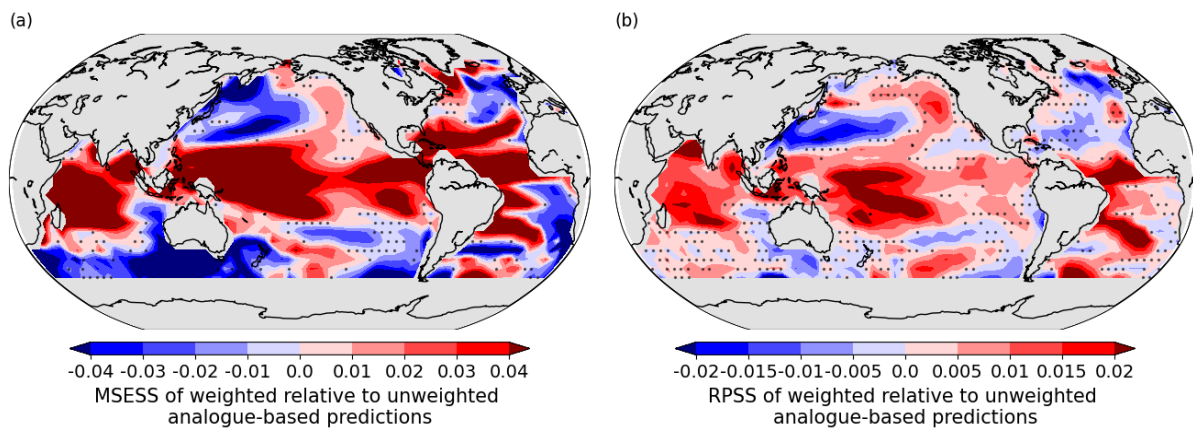


Figure 13: Pseudo-observation hindcasting: skill scores of the weighted ensemble relative to the unweighted ensemble. The skill metrics are (a) the mean squared error skill score (MESS), and (b) the ranked probability skill score (RPSS). MESS and RPSS are computed for SST average of hindcast years 1-10. Skill is aggregated for 316 pseudo-observations, with hindcast start years 1930-2015. Stippling indicates absence of significance when controlling for a false discovery rate of 10%. Figure adapted from Verjans et al. (2026).

We have used the calibrated performance-based weighting scheme to perform hindcasts of the real-world observational SST record. Here also, we have focused on decadal predictions, and we performed hindcasts over the 1930-2024 period. The performance-weighted ensemble shows large areas of skillful predictions, as evaluated with different evaluation metrics (Figure 14 a, c, e). When compared to the unweighted ensemble, we find regions of improved skill from performance-based weighting, most notably across the Pacific ocean, and in the southern Indian ocean (Figure 14 b, d, f). However, skill gains are not as pronounced compared to the pseudo-observation hindcasts (compare Figures 14 and 13). We have explained this apparent contradiction by three critical factors. First, there exist intrinsic predictability limits for decadal variability, and regions with low intrinsic predictability can necessarily only yield small gains in predictive skill. Second, the unweighted and weighted ensemble have a large similarity. Since the analogue selection needs to match the phase of internal climate variability, there is a large degree of overlap between the analogues selected in the weighted and unweighted ensembles. Third, there are large sampling uncertainties in the skill metrics given the short observational record (1930-2024), which only consists of ~9 independent decades. We verified this effect by sub-sampling the observational record, and showing that longer hindcast verification records systematically lead to larger skill gains of the performance-based weighted ensemble relative to the unweighted ensemble. This research is currently under review, with a preprint openly available (Verjans et al., 2026).

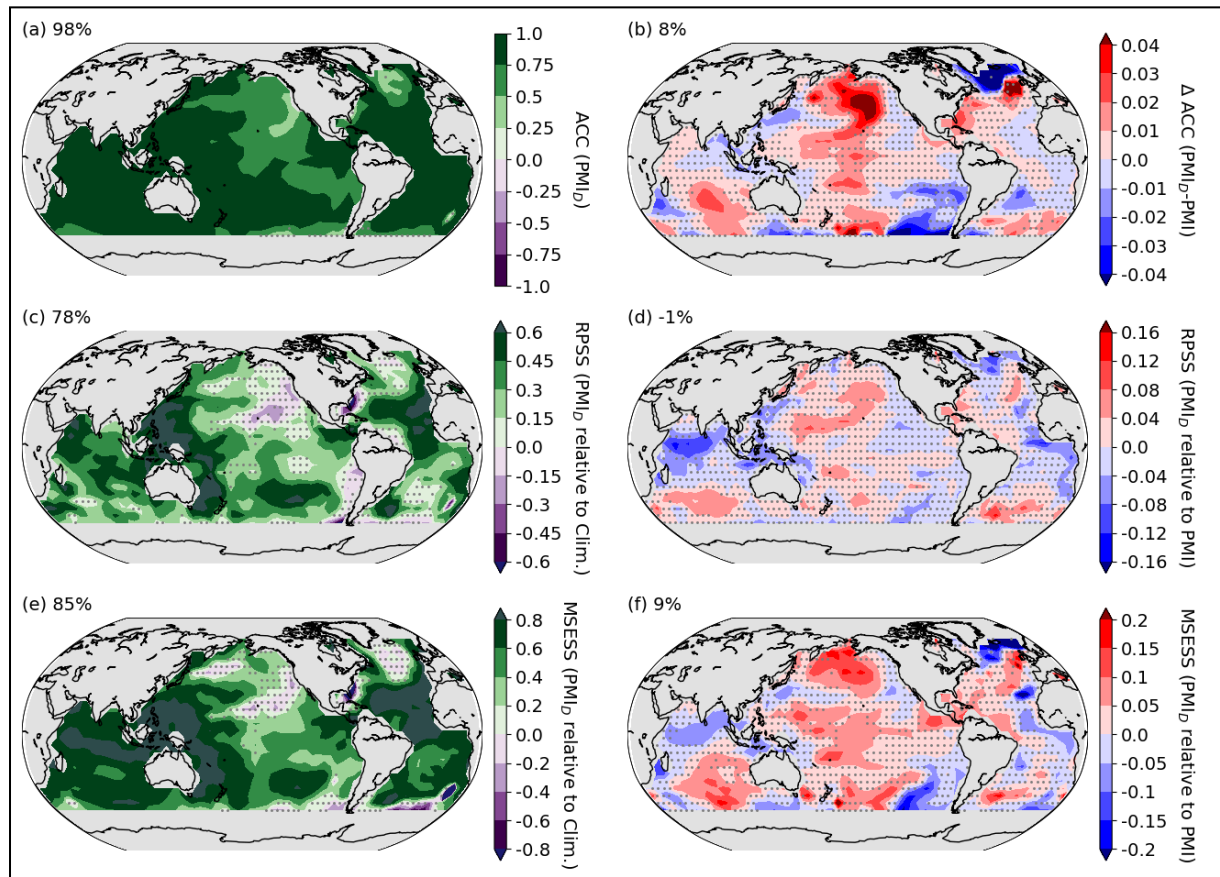


Figure 14: 1930-2024 hindcast performance. Evaluation of predicted SST average of hindcast years 1-10, verified against observations. (a) Anomaly correlation coefficient (ACC) of the performance-based weighted ensemble. (b) Difference in ACC between the weighted and unweighted ensembles. Ranked probability skill score (RPSS) of the performance-based weighted ensemble (c) relative to climatology and (d) relative to the unweighted ensemble as reference. Mean squared error skill score (MSESS) of the performance-based weighted ensemble (e) relative to climatology and (f) relative to the unweighted ensemble as reference. The total number of 10-year averaged hindcasts is 86, with 1930 and 2015 as first and last start dates. The net percentage of area with significant positive skill is given on top of each map. Stippling indicates absence of significance while controlling for a false discovery rate of 10%. Figure from Verjans et al. (2026).

5.4 Multi-variate extension of the analogue method using Earth observations

To date, the analogue-based prediction method described in the sections above has only used SST data for constraining initial states and weighting model performance. In this next phase of this project, we are extending this prediction framework to multiple variables, starting with sea-surface height (SSH). Spatiotemporal variability in SSH captures several aspects of climate dynamics. First, it

reflects changes in sub-surface ocean heat content, as temperature changes govern the density and therefore the volume of the water column. As such, SSH can potentially provide constraints more informative than SST on long prediction timescales, such as decadal to multi-decadal. Second, SSH is influenced by patterns of atmospheric pressure. Combining SST and SSH constraints can therefore inform on the state of the coupled ocean-atmosphere system. This coupling can improve predictive skill at shorter timescales (e.g., 1 to 5 years), since the chaotic behavior of the atmosphere likely erases initial state information beyond this timescale. In summary, combining SST and SSH constraints in the analogue selection method enables de facto constraining the coupled ocean-atmosphere state, and can improve predictive skill at different timescales due to the nature of information provided by SSH.

In this task, we exploit new generation satellite-based SSH observational products, such as the Copernicus sea-level anomaly record. This will enable a comparison of predictive skill in the periods post-satellite and pre-satellite. In the earlier period, SSH can only be determined through dynamical reanalyses rather than direct global observations.

In the early stages of this work, we have examined predictive skill in SST and SSH when combining these two variables in the constraints of the analogue selection. The balance between the SST and SSH constraints has been tested as two coefficients varying between 0 and 1 and summing to 1, i.e., constraining only on SST, only on SSH, or any intermediary between these two cases. In addition, we have tested the analogue-based selection using only the satellite SSH product from Copernicus, thus necessarily selecting analogues based only on the SSH constraint. Figure 15 shows how the skill of the predictions changes as the SST and SSH constraint coefficients change. SST predictions progressively degrade as the SST constraint coefficient decreases, suggesting they do not directly benefit from incorporation of SSH information at hindcast start time. In contrast, SSH skill rapidly increases, demonstrating that SSH predictions cannot rely exclusively on SST initial constraints. However, our findings also show that gain in SSH prediction skill saturates, and even starts decreasing as the SST constraint is increasingly ignored. This suggests that SST initial constraint is needed for optimal SSH predictions. The skill evaluated against the satellite SSH product (dot in Figure 15) shows a limitation of these products as benchmarks for multi-year hindcasts. Due to the short record availability, statistical significance of skill is hard to establish. Such products may therefore be used as confirmation of skill evaluated in longer-term reanalysis products, rather than a unique validation data set.

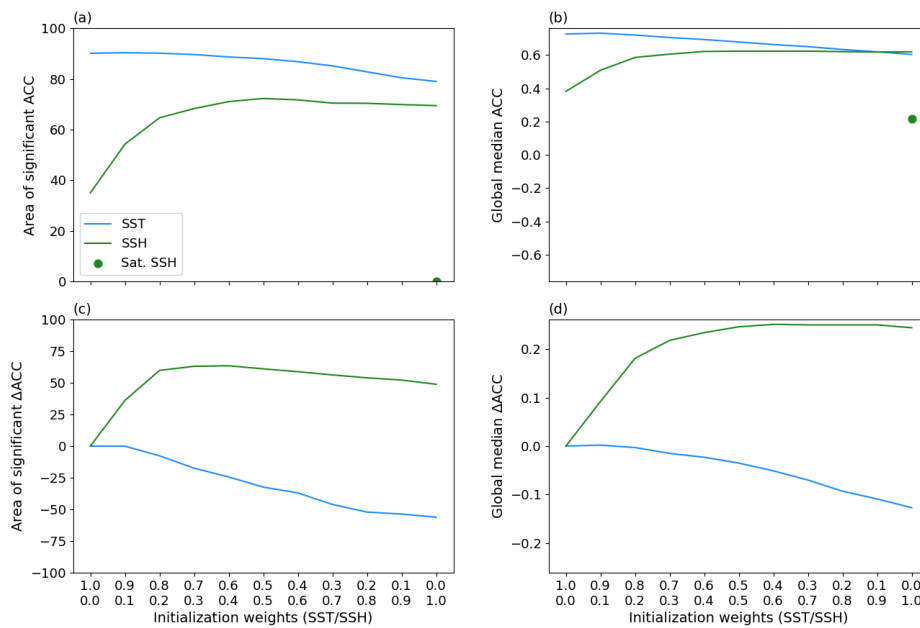


Figure 15: Hindcast prediction skill for SST (blue curves) and SSH (green curves) as a function of weights for initialization constraints varying individually between 0 and 1. Skill measures are (a, b) the anomaly correlation coefficient (ACC), and (c, d) the difference in ACC relative to the SST-only constraint (Δ ACC). Sub-panels (a, c) show the percentage global area of significance and (b, d) the global median. Skill is evaluated for predictions average of hindcast years 1 to 3, and evaluated against data from reanalysis over the period 1958-2014. Also shown is the skill for predicting the satellite-based SSH observations (green dot), which necessarily uses an SSH-only initial constraint, and for which hindcast skill is substantially lower due to the short verification record available (1993-2014).

Figure 16 shows a spatial analysis of these prediction skill evaluations, illustrating skill maps for the SST-only, SSH-only, and equally-weighted SST and SSH initialization constraints. The maps show that increasing the SSH initialization constraint weight degrades SST prediction skill most in the North-Pacific and South-Atlantic oceans, and increases SSH prediction skill most in the tropical oceans. However, even for SSH predictions, skill is degraded in the East-Pacific if the SST initialization constraint is neglected. Here also, the validation based on the satellite product suffers from absence of significance. Still, we identify similar skill patterns with the validation based on the reanalysis. For example, relatively higher levels of skill in the West and Central Pacific, in the East Atlantic, and in the South-West Indian ocean (Fig. 16 h, k).

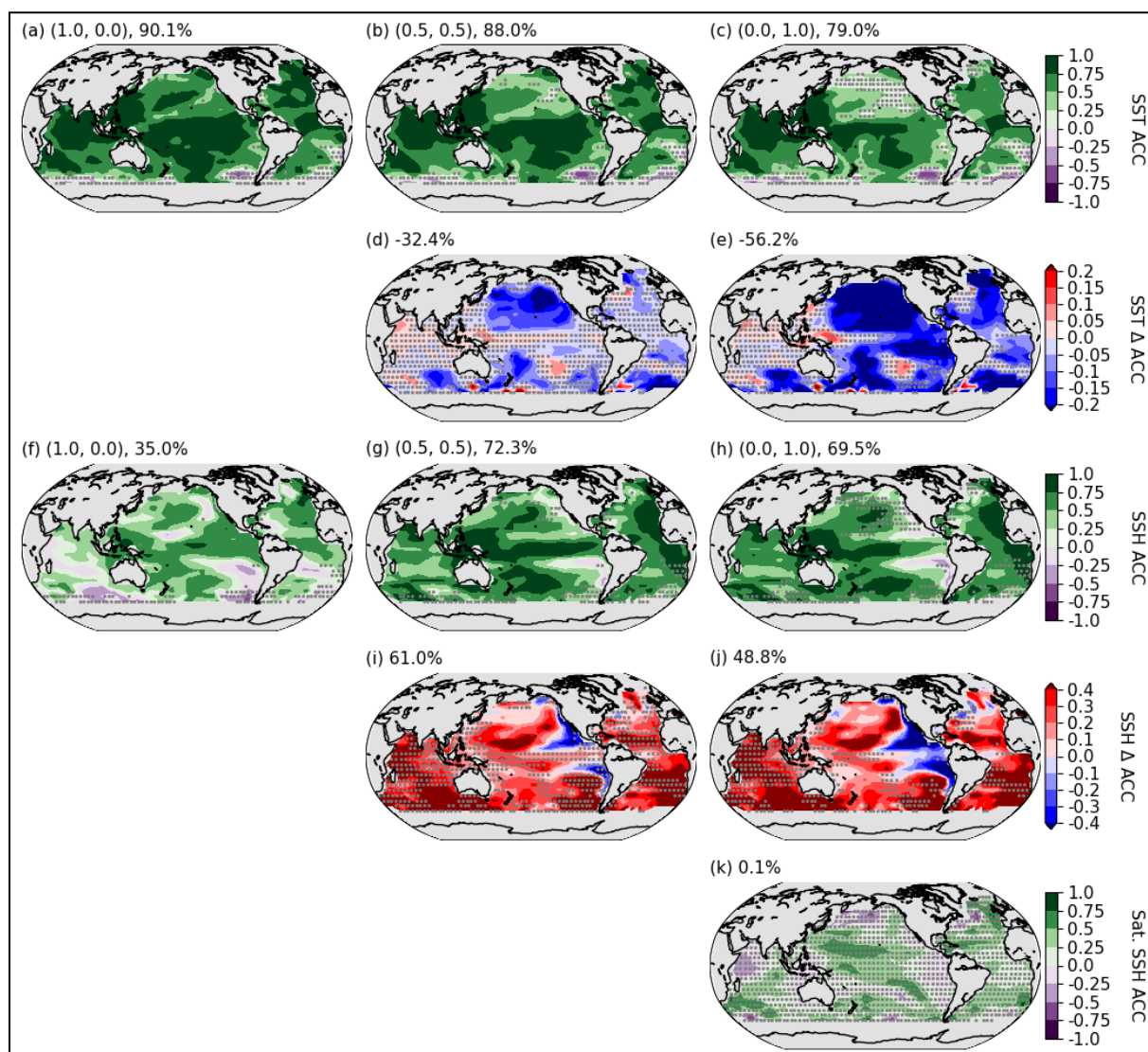


Figure 16: Hindcast prediction skill for SST (first two rows) and SSH (last three rows) for different configurations of the initialization constraints: constraining only on SST (left column), only on SSH (right column) and equally on SST and SSH (middle column). Skill measures are (a, b, c, f, g, h, k) the anomaly correlation coefficient (ACC), and (d, e, i, j) the difference in ACC relative to the SST-only constraint (Δ ACC). Skill is evaluated for predictions average of hindcast years 1 to 3, and evaluated against reanalysis data over the period 1958-2014. Also shown (k) is the skill for the satellite-based SSH observations, which necessarily uses an SSH-only initial constraint, and for which hindcast skill is substantially lower due to the short verification record available (1993-2014). The net percentage of area with significant positive skill is given on top of each map, as well as the SST and SSH initialization constraint weights. Stippling indicates absence of significance while controlling for a false discovery rate of 10%.

There are several avenues to bring this preliminary work forward. First, we will extend the hindcast period beyond 2014 and up to present-day once the necessary model

data will be processed. Second, we will analyze more specifically differences in prediction skill pre- versus post-1993, to evaluate the influence of the incorporation of satellite-based SSH measurements on the quality of the analogue selection and the prediction evaluation. Third, we will analyze under which specific conditions the SSH initial constraint brings added skill in predictions, as this is expected to be state-dependent, where different modes of climate variability imply differences in SST-SSH coupling. To this end, we will use the pseudo-observational framework to enable analysis of multiple hundreds of model realizations – necessary to ensure statistical robustness. Finally, we will extend the model evaluation deviance metric (see previous subsection 2) to the multi-variate SST-SSH system, hence evaluating model representation of coupled ocean-atmosphere dynamics. This will also extend our approach to performance-based multi-variate analogue predictions.

6. ML extension of high-resolution reanalysis

6.1 Introduction

Reanalyses play an important role in understanding climate change and to mitigate its impact. The ERA5 global reanalysis (Hersbach et al., 2020) has a comprehensive spatial and temporal extent. However, its resolution is limited (approx 25km) and physically valid higher resolution data has the potential to add value to important applications. We explore the extension of local high-resolution data, either in space or time, by tethering it to a coarser dataset. For this, we learn the correlation between, for example ERA5 and CERRA (a local high-resolution reanalysis for Europe). At inference time, ERA5 data can then be used to generate CERRA-like output when- or wherever ERA5 is available. We demonstrate this by generating a back-extension of CERRA. We explore a number of different training strategies for this task, including a super-resolution-type approach and forecasting/dynamic downscaling. We also investigate to what extent pre-training on ERA5 can improve the predictions. The work will be expanded in Task 2.2 where climate downscaling is implemented, which can as well be seen as extending one dataset to a higher resolution one. For this, we want to further refine the training strategies with self-supervised approaches to better learn the correlations between the datasets, potentially further leveraging observational data beyond its assimilation in the parent reanalysis datasets.

6.2 Methods

To learn the correlation between global and regional reanalyses, the ECMWF team built on the WeatherGenerator model whose architecture is shown in Figure 17. It is a highly multi-modal transformer-based encoder-processor-decoder architecture that can ingest and predict a wide range of structured and unstructured data. The encoder maps all points of the physical input fields to a latent representation of fixed dimension, which is spatially structured. Depending on the training objective, the processor applies a learned function to the latent state, e.g. for time stepping. Finally, the decoder maps the latent vector back to the physical fields. For learning the correlation between ERA5 and CERRA, we follow a two-stage training schedule. In the first stage, we train a forecast model for the ERA5 global reanalysis on an approx. 1-degree grid. To better constrain the forecast objective, training is performed on three forecast steps using latent space roll-out, i.e. the processor is called autoregressively. In the second step, the pretrained encoder and processor are used together with a new decoder for the high-resolution CERRA data. While the weights of the encoder are frozen during this stage, the processor and the CERRA-specific decoder are fine-tuned. The model is trained to predict all CERRA variables. High resolution geoinformation channels of orography and land-sea-mask from the CERRA dataset are used as additional information for the decoder.

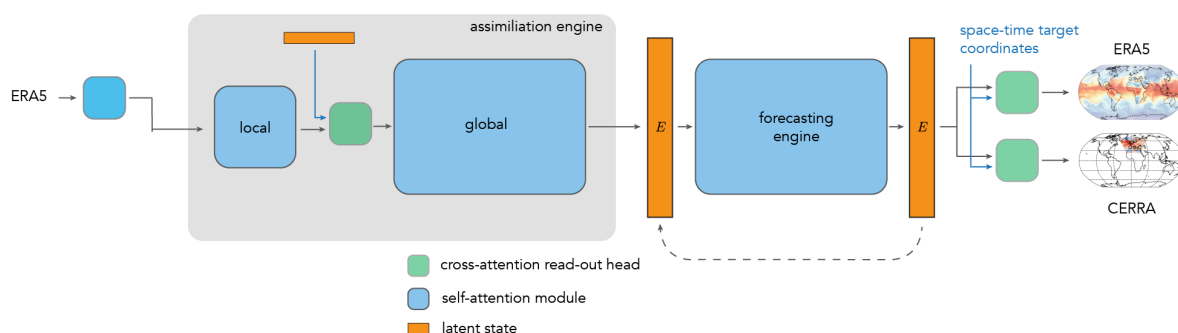


Figure 17: Architecture of the WeatherGenerator model.

6.3 Results

We train the model on a training period from 2000 to 2023 and evaluate the model's backfilling capabilities in the year 1999. The evaluation is performed on a total of 240 samples across the entire year, taken from the first 20 timesteps in each month (to sample the full annual cycle). As a baseline, the ERA5 at approximately 1 deg is interpolated onto the CERRA grid using nearest neighbors interpolation. We evaluate

the model on two near-surface fields, i.e. u-component of wind and temperature. Skill scores are shown in Table 5. While the interpolation is doing well for smoother fields like temperature, the advantages of the ML model becomes more pronounced for more complex fields like wind.

Variable	Method	RMSE	Pixel-wise RMSE	Pixel-wise RMSE STD
1000 hPa wind u-component	Baseline	2.54 m/s	2.32 m/s	1.02 m/s
	WeatherGen	1.40 m/s	1.34 m/s	0.41 m/s
2m temperature	Baseline	1.96 K	1.55 K	1.18 K
	WeatherGen	1.68 K	1.56 K	0.61 K

Table 5: Performance scores comparing the model predictions with the baseline from the nearest neighbor interpolation calculated from a total of 240 samples from year 1999, i.e. the 20 first timesteps from each month.

Maps for single sample examples for wind and temperature are shown in Figure 18 and 19, respectively. The plots are in line with the quantitative scores observed in Table 5. The advantages of the trained model are apparent in the error of the wind field as well as in areas of rough orography such as over the Alps. The interpolation shows good results mainly over smooth regions.

Comparison: ERA5 vs Model vs CERRA - U_1000
Europe Region | 1999-10-01T06:00:00

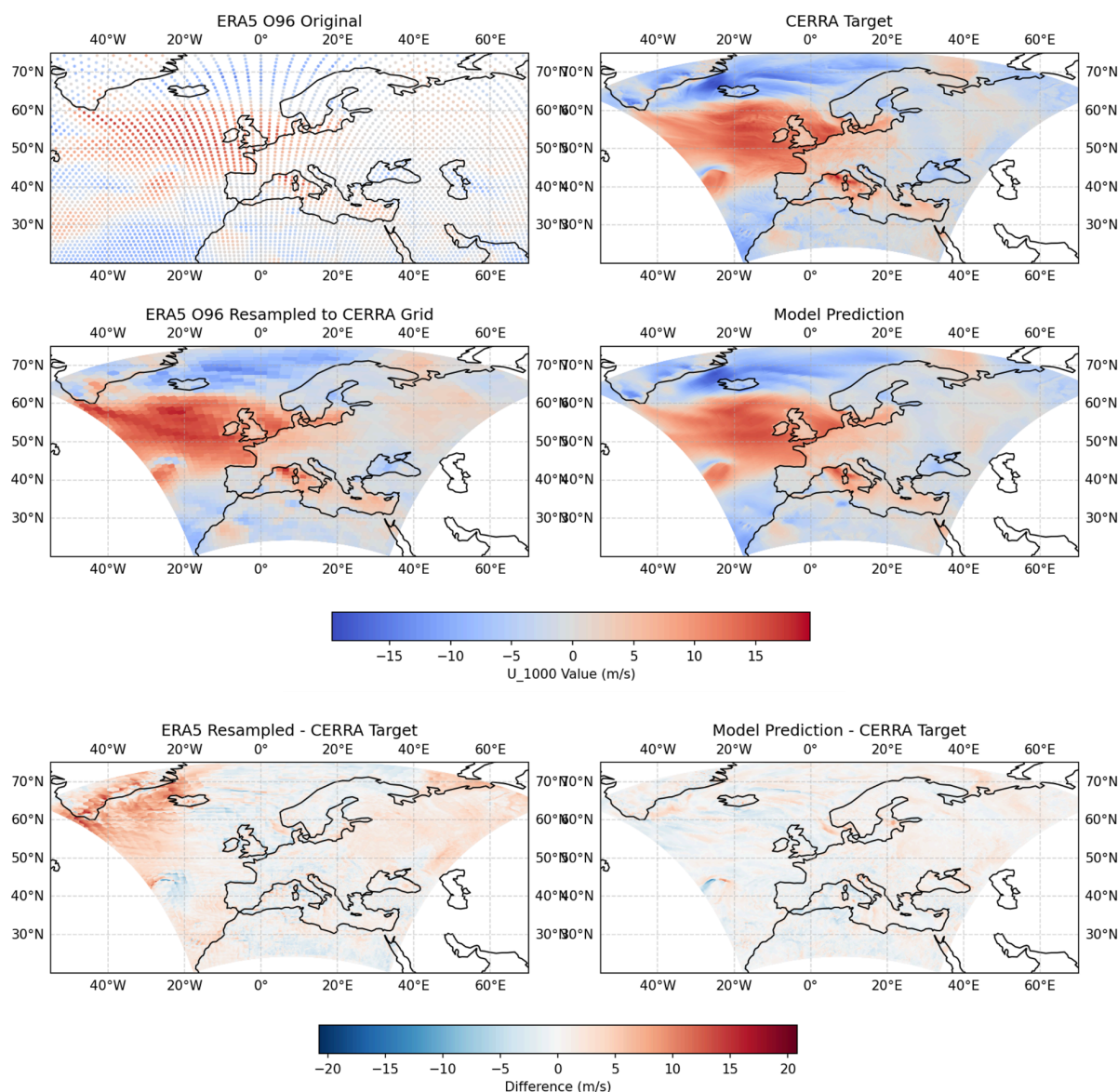


Figure18: Maps for 1000 hPa u-component wind for a single sample on 1999-10-01 6:00. Top row, left: global ERA5 O96 input field; top row, right: regional CERRA target field; second row, left: ERA5 O96 interpolated to CERRA grid (nearest neighbor); second row, right: model prediction; third row, left: difference between interpolated ERA5 and target CERRA; third row, right: difference between model prediction and target CERRA.

Comparison: ERA5 vs Model vs CERRA - 2T
Europe Region | 1999-10-01T06:00:00

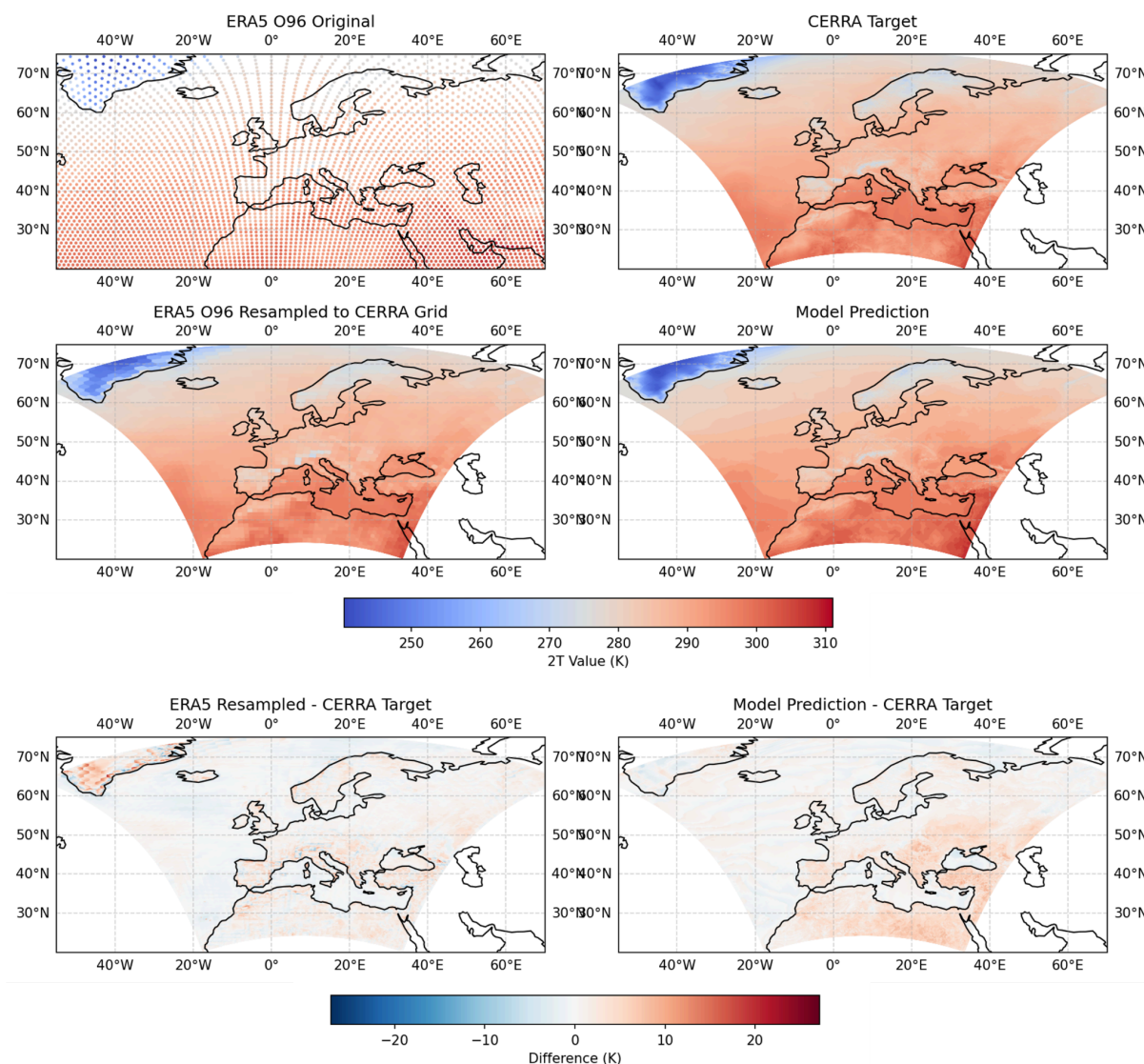


Figure 19: Maps for 2m temperature for 1999-10-01 6:00. Top left: Global ERA5 O96 input field; top right: Regional CERRA target field;; second row, left: interpolated ERA5 O96; second row, right: model prediction; bottom left: Difference between interpolated ERA5 and target CERRA; bottom right: Difference between model prediction and target CERRA.

7. Conclusions

7.1 Key results and impact for EXPECT

This deliverable provides a central contribution to **Theme 1: Data for the generation of new climate knowledge**, whose objective is to combine EO, climate simulations and ML to overcome key limitations in understanding and predicting regional climate variability and extremes. By developing new methods to bridge observational and model data, D1.1 directly supports the overarching EXPECT goal of establishing a prototype capability for integrated attribution and prediction of regional climate change.

The main results of this deliverable are:

(1) New AI-based reconstructions of observational climate datasets

DL methods have been developed and applied to reconstruct incomplete records of temperature extremes and precipitation. These products provide spatially and temporally consistent fields that preserve observed variability while substantially reducing interpolation artefacts. Particular emphasis is placed on ensuring that the reconstructions remain physically consistent across variables, regions, and timescales, so that the resulting fields are not only statistically plausible but also climatologically meaningful. The DL reconstructions extend the usable observational record into periods and regions with sparse coverage, thereby enabling more robust detection of long-term changes and improved process-based model evaluation. This directly addresses one of the key knowledge gaps identified by EXPECT: the limited availability of consistent observations for analysing regional climate change on inter-annual to multi-decadal timescales.

(2) Observation-constrained initialization for decadal prediction

The hybrid ML data-assimilation framework demonstrates that transferring physically consistent ocean information from well-observed to poorly observed periods improves the representation of North Atlantic variability and enhances prediction skill for the Subpolar Gyre. This provides a proof of concept that observational constraints derived with ML can improve initialized climate predictions in data-sparse periods without degrading skill elsewhere. This capability is essential for the EXPECT objective of delivering more reliable and physically interpretable near-term climate predictions.

(3) EO-based diagnostics of land–atmosphere coupling regimes

The identification of hydrological regimes from satellite observations provides

physically interpretable, observation-based benchmarks for evaluating land-surface processes in climate models. These diagnostics enable a process-oriented assessment of model performance and help constrain one of the major sources of uncertainty in regional climate projections, namely land-atmosphere interactions and their role in extremes.

(4) Analogue-based prediction with observational and performance constraints

The development of performance-weighted and multivariate analogue methods demonstrates a computationally efficient pathway to constrain internal climate variability and improve near-term predictive skill. Within D1.1, this approach complements dynamical prediction by providing an alternative, data-driven route to near-term climate information, with different strengths and limitations. Whereas dynamical prediction explicitly represents the evolving climate system through initialized model simulations, analogue-based methods draw predictive information from observed and simulated precedents, offering a useful benchmark and a flexible framework for testing observational and performance constraints. As such, they offer complementary perspectives on predictability, with dynamical models exploiting process-based memory in the climate system and analogue methods exploiting recurring predictable patterns in the historical record and free-running model simulations. By combining observational constraints with objective model-performance metrics, this work contributes to the EXPECT ambition of linking physical understanding, model evaluation and prediction within a consistent framework.

(5) Machine-learning extension of high-resolution reanalysis information

The WeatherGenerator-based framework shows that relationships between global and regional reanalyses can be learned and used to generate high-resolution information beyond the native temporal coverage of regional products. To maximise scientific value, these extensions should not only be statistically skillful but also physically consistent, preserving meaningful relationships across scales and variables. This strengthens confidence that the generated high-resolution fields are suitable for process-based interpretation as well as data-efficient analysis. This provides a scalable strategy for exploiting recent investments in high-resolution climate datasets and supports the project-wide effort to efficiently analyse large and heterogeneous data volumes.

Contribution to the integrated EXPECT capability

Taken together, the developments presented in this deliverable establish the methodological foundation for integrating EO and ESMs across four complementary dimensions:

- reconstruction of incomplete observations,
- process-based model evaluation,
- observation-constrained prediction initialization,
- extension and harmonisation of multi-resolution datasets.

This integration is a prerequisite for explaining and credibly predicting regional climate change, because it enables the consistent quantification of the processes that drive observed and simulated variability. In this way, D1.1 provides essential building blocks for the subsequent attribution, prediction and hazard-assessment activities in EXPECT Themes 2 and 3.

Wider scientific and societal relevance

The methodological and data developments achieved in this deliverable strengthen the scientific basis for reliable regional climate information, which is a prerequisite for effective climate-risk assessment and adaptation planning in Europe. By extending observational records, improving the representation of key processes in models, and enhancing the initialization of near-term predictions, this work contributes to reducing fundamental uncertainties that currently limit the usability of climate information for decision-making on timescales from years to decades.

The new AI-based reconstruction approaches that exploit a broader set of input data, including ESM and high-resolution reanalysis data enable the exploitation of sparse historical observations and therefore improve the assessment of long-term changes in climate extremes and regional variability. This is particularly relevant for sectors that require robust historical baselines—such as water management, agriculture, energy systems and infrastructure planning—where risk estimates depend critically on the accurate characterization of past variability and trends.

The observation-constrained prediction approaches developed here contribute to more physically consistent and more skilful near-term climate predictions. Such predictions are key for Climate Service organizations to provide timely and actionable information for adaptation planning, investment cycles and policy implementation.

In addition, the scalable ML workflows for combining EO, reanalyses and climate simulations directly support ongoing European efforts to exploit large and heterogeneous climate data volumes, including the Copernicus Climate Change Service and Destination Earth. By enabling efficient use of high-resolution datasets and by providing transferable, computationally efficient methods, the results help maximise the value of existing European investments in climate observations and modelling.

Finally, the process-based evaluation diagnostics and performance-weighted prediction frameworks contribute to building trust in climate information, as they link predictive skill to physically interpretable mechanisms and observational constraints. This is essential for the uptake of climate information by users and for the development of transparent, traceable climate assessments in line with the objectives of the WCRP Lighthouse Activity on Explaining and Predicting Earth System Change and the broader European climate-services landscape.

7.2 Contribution to next WPs and long-term project objectives

The developments achieved in Task 1.1 provide methodological and data foundations that will be further exploited across several EXPECT work packages. The reconstructed observational datasets of temperature extremes and precipitation provide spatially complete observational constraints that are relevant in WP6 Tasks 6.1–6.2 for downstream analyses of European temperature and precipitation extremes and associated drivers of regional climate variability. The hybrid ML data assimilation framework established in Task 1.1 will be further advanced in WP2 Task 2.3 through its application to MPI-ESM hindcast experiments and evaluation of forecast skill, biases, and uncertainties, including targeted analysis of European summer extremes and North Atlantic drivers. The analogue-based initialization approach developed in Task 1.1 is expected to complement WP4 Tasks 4.1–4.2 by enabling observation-constrained ensemble sub-selection and initialization for integrated attribution–prediction and calibrated decadal prediction analyses. The EO-based diagnostics of land-surface temperature and soil-moisture coupling developed in Task 1.1 can inform WP5 Task 5.2 by providing observational characterisation of land–atmosphere interaction regimes and water-limited conditions relevant for analyses of summer temperature and precipitation extremes. Finally, the extended high-resolution reanalysis fields derived from ERA5–CERRA provide spatially consistent regional climate information that can support regional

analyses and downscaling activities across WP2–WP4. Together, these contributions propagate EO-constrained data, diagnostics, and initialization capabilities from WP1 into attribution, prediction, and application-oriented components of EXPECT.

EXPECT Data availability

Section	Description	Link/DOI
Section 1	Annual update of Climate Reconstruction AI (CRAI) infilled HadCRUT5 of near-surface temperature change 1850 to 2024	https://doi.org/10.5281/zenodo.15622090
Section 1	HadEX-CAM dataset: original and deep learning infilled TX90p, TN90p, TX10p, TN10p ETCCDI Indices	https://doi.org/10.26050/WDCC/HadEX-CAM

EXPECT Code repositories

Section	Description	Link/DOI
Section 1	Software to train/evaluate models to reconstruct missing values in climate data based on a U-Net with partial convolutions.	https://github.com/FREVA-CLINT/climatereconstructionAI
Section 5	Code to perform performance-based analogue predictions, example for 1930–2024 sea-surface temperature decadal predictions	https://doi.org/10.5281/zenodo.18033875

EXPECT publications

Beobide-Arsuaga, G., Suarez-Gutierrez, L., Barkhordarian, A. et al. (2025). Increasing central and northern European summer heatwave intensity due to forced changes in internal variability. *Nat Commun* 16, 9485. <https://doi.org/10.1038/s41467-025-65392-w>

Donat, M.G., Mahmood, R., Cos, P., Ortega, P., Doblas-Reyes, F. (2024). Improving the forecast quality of near-term climate projections by constraining internal variability based on decadal predictions and observations. *Environmental Research: Climate* 3(3), 035013 <https://doi.org/10.1088/2752-5295/ad5463>

Forster, P. M., Smith, C., Walsh, T., Lamb, W. F., Lamboll, R., Cassou, C., Hauser, M., Hausfather, Z., Lee, J.-Y., Palmer, M. D., von Schuckmann, K., Slangen, A. B. A., Szopa, S., Trewin, B., Yun, J., Gillett, N. P., Jenkins, S., Matthews, H. D., Raghavan, K., Ribes, A., Rogelj, J., Rosen, D., Zhang, X., Allen, M., Aleluia Reis, L., Andrew, R. M., Betts, R. A., Borger, A., Broersma, J. A., Burgess, S. N., Cheng, L., et al. (2025). Indicators of Global Climate Change 2024: annual update of key indicators of the state of the climate system and human influence, *Earth Syst. Sci. Data*, 17, 2641–2680, <https://doi.org/10.5194/essd-17-2641-2025>

Pléziat, É., Dunn, R.J.H., Donat, M.G. and Kadow C. (2024). Artificial intelligence reveals past climate extremes by reconstructing historical records. *Nat Commun* 15, 9191), <https://doi.org/10.1038/s41467-024-53464-2>

Satpathy, S.S., Franzke, C.L.E., Verjans, V., and O’Kane, T. (2026). Anthropogenic climate change leads to a pronounced reorganisation of wintertime North Atlantic atmospheric circulation regimes. *Commun Earth Environ* 7, 155, <https://doi.org/10.1038/s43247-026-03180-0>

Thorne, P. W., Nicklas, J. M., Kennedy, J. J., Calvert, B., et al. (2026). How well can we quantify when 1.5 °C of global warming has been exceeded?, *Earth Syst. Sci. Data Discuss.* [preprint], <https://doi.org/10.5194/essd-2025-825>, in review.

Verjans, V., Donat, M., Delgado-Torres, M., and DelSole, T. (2026). Large potential of performance-based model weighting to improve decadal climate forecast skill, *npj Climate and Atmospheric Science* [preprint], <https://doi.org/10.21203/rs.3.rs-8442113/v1>, in review.

References

- Brune, S., & Baehr, J. (2020). Preserving the coupled atmosphere–ocean feedback in initializations of decadal climate predictions. *Wiley Interdisciplinary Reviews: Climate Change*, 11(3), e637.
- Copernicus Climate Change Service (C3S) Climate Data Store (CDS). DOI: 10.24381/cds.d7782f18
- Dunn, R. J., & Coauthors (2020). Development of an Updated Global Land In Situ-Based Data Set of Temperature and Precipitation Extremes: HadEX3. *Journal of Geophysical Research: Atmospheres*, 125 (16), 1–28
- Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E. (2016). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization, *Geosci. Model Dev.*, 9, 1937–1958, <https://doi.org/10.5194/gmd-9-1937-2016>
- Feldman, A. F., Short Gianotti, D. J., Trigo, I. F., Salvucci, G. D., and Entekhabi, D. (2019). Satellite-based assessment of land surface energy partitioning–soil moisture relationships and effects of confounding variables. *Water Resources Research*, 55(12), 10657–10677.
- Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., ... and Thépaut, J. N. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730), 1999–2049
- Ho, J., Jain, A. & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* 33, 6840–6851.
- Kadow, C., Hall, D. M., and Ulbrich, U. (2020). Artificial intelligence reconstructs missing climate information. *Nature Geoscience*, 13 (6), 408–413. <https://doi.org/10.1038/s41561-020-0582-5>
- Lam, R., Sanchez-Gonzalez, A., Willson, M., et al. (2023). Learning skillful medium-range global weather forecasting. *Science*, 382, 1416–1421. <https://doi.org/10.1126/science.adi2336>
- Lentz, S., Brune, S., Kadow, C., and Baehr, J. (2025). Improving ocean reanalyses of observationally sparse regions with transfer learning. *Scientific Reports*, 15(1), 2640.

- Liu, G., Reda, F.A., Shih, K. J., Wang, T. C., Tao, A., and Catanzaro, B. (2018). Image Inpainting for Irregular Holes Using Partial Convolutions. Computer Vision – ECCV 2018. ECCV 2018. Lecture Notes in Computer Science, 11215. Springer, https://doi.org/10.1007/978-3-030-01252-6_6
- Mahmood, R., Donat, M., Ortega, P., Doblas-Reyes, F., Delgado-Torres, C., Samsó, M., and Bretonnière, P. (2022). Constraining low-frequency variability in climate projections to predict climate on decadal to multi-decadal timescales—a poor man's initialized prediction system. Earth System Dynamics 13 (4), 1437-1450. <https://doi.org/10.5194/esd-13-1437-2022>
- Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. arXiv, <https://doi.org/10.1109/ACCESS.2021.3053408, 1505.04597>
- Schneider, U. et al., (2022) DOI: 10.5676/DWD_GPCC/FD_M_V2022_100
- Shi, Y., Huang, Z., Feng, S., Zhong, H., Wang, W., & Sun, Y. (2020). Masked Label Prediction: Unified Message Passing Model for Semi-Supervised Classification. arXiv:2009.03509.
- Trigo, I. F., DaCamara, C. C., Viterbo, P., Roujean, J.-L., Olesen, F., et al. (2011). The Satellite Application Facility on Land Surface Analysis. Int. J. Remote Sens., 32, 2725-2744, doi: 10.1080/01431161003743199.