CopERnIcus climate change Service Evolution

D7.4 Time varying Lake cover, Land cover and LAI and extension of CONFESS vegetation data back to 1925 datasets

Due date of deliverable	30.04.2025
Submission date	30.05.2025
File Name	CERISE-D7-4-V1.0
Work Package /Task	WP7
Organisation Responsible of Deliverable	Barcelona Supercomputing Center (BSC)
Author name(s)	Etienne Tourigny, Amirpasha Mozaffari, Vinayak Huggannavar, Iria Ayan, ,Margarita Choulga, Souhail Boussetta, David Fairbairn
Revision number	1.0
Status	Issued
Dissemination Level	Public (PU)

Funded by the European Union

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

Table of Contents

1 Introduction

The CERISE project has outlined requirements for the boundary conditions of ecLand, the land surface component of the ECMWF Integrated Forecasting System (IFS), which will be used for generating the second CERISE high-resolution land reanalysis prototype (ERA6-land-Pv2) from the present back to the year 1925. This prototype relies on the generation of high-resolution monthly varying lake cover, and time-varying dataset of land cover and LAI (Leaf Area Index). This report presents the CERISE time varying vegetation and lake datasets.

Many reanalysis systems have historically relied on fixed or climatological land-cover representations – for example, ERA5-Land prescribes a single, seasonally averaged vegetation cycle throughout its entire archival period (Muñoz-Sabater et al., 2021). However, recent advances now allow the use of time-varying surface inputs: ESA's (European Space Agency) CCI (Climate Change Initiative) Land Cover project delivers annual, 300 m global maps of discrete land-cover types from 1992 to the present (Li et al., 2018), while the LUH2 (Land Use Harmonization 2) database provides 0.25° grids of fractional land-use states and transitions extending back to 850 CE (Chini et al., 2021). Despite their complementary strengths, these datasets differ in spatial resolution, categorical schemes, and temporal extent, limiting their direct use in next-generation reanalysis, carbon-cycle and hydrological models that demand consistent, high-resolution land-cover forcing. Building on recent efforts to reconcile multi-source land-surface information, we apply a state-of-the-art harmonization framework to merge LUH2 fractional trajectories with ESA CCI's discrete classes, producing an annual, 1 km-resolution time series of land-cover maps aligned to the CCI classification system. This harmonized product fills a critical gap in both spatial detail and temporal continuity, enabling more realistic simulations of land–atmosphere interactions and better attribution of terrestrial change over the past millennium.

Lakes modify the structure of the atmospheric boundary layer. They can have a significant impact on local climate (over 1°K difference in 2-meter temperature (Samuelsson et al., 2010) and on local weather (up to 10°K difference in 2-meter temperature (see Eerola et al., 2014)). At the European Centre for Medium-Range Weather Forecasts (ECMWF), lake parametrization was introduced in 2015. Inland water bodies (i.e. lakes, reservoirs, rivers and coastal waters) are simulated by the Fresh-water Lake model Flake. The IFS model is used for global weather forecast production from medium to seasonal range, and for reanalysis (e.g. ERA5) generation. It has been shown in previous studies that monthly varying lake mask has a significant positive impact on regions with prolonged rain and dry seasons, for example in Malaysia, Indonesia and Papua New Guinea (Kimpson et al., 2023).However, the current lake mask used in IFS is still constant over time and represents permanent water over the 34-year period (i.e. 1984-2018).

The LAI is one of the most critical variables governing land-atmosphere exchange (Fang et al., 2019). Physically, LAI controls how the land surface partitions energy and water, through plant transpiration via root extraction of soil moisture, and evaporation of intercepted rain. It also controls interception of sunlight, therefore altering the albedo and surface heating (Boussetta et al., 2015). Thus, a realistic and dynamic LAI is expected to improve the prediction of temperature and relative humidity, especially in extreme events such as heatwaves and droughts (Duveiller et al., 2022). In addition, a time-varying LAI is expected to enhance climate predictability at multiple scales (Alessandri et al., 2017).

Together, time-varying land cover, lake cover, and LAI provide a more realistic representation of land surface processes, which is essential for improving the accuracy of climate reanalysis and forecasts across temporal and spatial scales.

1.1 Background

The scope of CERISE is to enhance the quality of the Copernicus Climate Change Service (C3S) reanalysis and seasonal forecast portfolio, with a focus on land-atmosphere coupling. It will support the evolution of C3S, over the project's four-year timescale and beyond, by improving the C3S climate reanalysis and the seasonal prediction systems and products towards enhanced integrity and coherence of the C3S Earth system Essential Climate Variables. CERISE will develop new and innovative ensemble-based coupled land-atmosphere data assimilation approaches and land surface initialisation techniques to pave the way for the next generations of the C3S reanalysis and seasonal prediction systems. These developments will be combined with innovative work on observation operator developments integrating Artificial Intelligence (AI) to ensure optimal data fusion fully integrated in coupled assimilation systems. They will drastically enhance the exploitation of past, current, and future Earth system observations over land surfaces, including from the Copernicus Sentinels and from the European Space Agency (ESA) Earth Explorer missions, moving towards an all-sky and all-surface approach. For example, land observations can simultaneously improve the representation and prediction of land and atmosphere and provide additional benefits through the coupling feedback mechanisms. Using an ensemble-based approach will improve uncertainty estimates over land and lowest atmospheric levels. By improving coupled land-atmosphere assimilation methods, land surface evolution, and satellite data exploitation, R&I inputs from CERISE will improve the representation of long-term trends and regional extremes in the C3S reanalysis and seasonal prediction systems. In addition, CERISE will provide the proof of concept to demonstrate the feasibility of the integration of the developed approaches in the core C3S (operational Service), with the delivery of reanalysis prototype datasets (demonstrated in pre-operational environment), and seasonal prediction demonstrator datasets (demonstrated in relevant environment). CERISE will improve the quality and consistency of the C3S reanalysis systems and of the components of the seasonal prediction multi-system, directly addressing the evolving user needs for improved and more consistent C3S Earth system products.

1.2 Scope of this deliverable

1.2.1 Objectives of this deliverables

The objective of this is to document the methodology and results of the three datasets developed in CERISE for:

- Time-varying Land Cover
- Time-varying Leaf Area Index (LAI)
- Time-varying lake cover

1.2.2 Work performed in this deliverable

In this deliverable the work outlined in The Description of Action reads as such:

"Task 7.3: Create and assess a consistent extension of CONFESS vegetation data back to 1925 at nominal 1km resolution for reanalysis."

"Create and assess time-varying Land Cover dataset (CERISE-LC) from 1993 back to 1925: CCI-LC product will be remapped to 1km resolution by modal aggregation and an established methodology will be used to extend it back to 1925 using LUH2 historical land-use data (Hurtt et al., 2020) and time-varying lake cover from T7.4. An evaluation of the dataset will be performed with existing regional land cover classifications based on high-resolution satellite imagery before 1993."

CERISE

The CERISE framework ensures spatial and temporal consistency between land use/land cover (LULC) and lake datasets by applying harmonized methodologies across both products. All datasets are generated at a nominal 1 km horizontal resolution and cover the historical period from 1925 to 2021, enabling seamless integration into reanalysis and modeling workflows. For LULC, the CERISE-LC dataset is derived by remapping the ESA CCI-LC product to 1 km resolution using modal aggregation. The time series is extended back to 1925 using a transition-based approach informed by LUH2 historical land-use data (Hurtt et al., 2020). This method preserves consistency in land-use dynamics while ensuring continuity in land cover classification over time. Additionally, the reconstruction incorporates time-varying lake cover information from Task 7.4 to ensure inland water bodies are treated consistently across all land cover classes. Some key points achieved in this task are as follows:

- LUH2 and ESA-CCI land-use datasets were harmonized to a common spatial resolution and unified land-use classification using a crosswalk table (CwT).
- The modeling framework operates on LUH2 subgrids, each representing approximately 28 × 28 block of ESA-CCI grid cells.
- Annual land-use transitions were simulated using gross transition values from LUH2.
- LUH2 transitions were reclassified to match the decided CwT to generate harmonized land-use classes before application.
- A 6 × 6 transition matrix was constructed for each subgrid and year, representing transitions among six key land-use classes: forest, shrubland, cropland, pasture, urban, and barren.
- Transitions were initially assigned randomly within each land-use mask.The final implementation uses controlled random assignment, adaptable to simulation needs.
- Inconsistencies were identified in pasture transitions due to mismatches in baseline forest and pasture extents between LUH2 and ESA-CCI.
- When insufficient pasture area was available, the model redirected pasture-to-natural transitions via cropland, but only after first synthesizing the required pasture area.

“Create and assess time-varying LAI dataset (CERISE-LAI) from 1993 back to 1925: produce a time-varying LAI dataset based on the CONFESS-LAI product and consistent with the Land Cover classification above. An independent evaluation will be performed with the CONFESS-LAI product from 1982-1993 based on AVHRR-GEOV2.”

In CERISE, time-varying LAI datasets were created based on ML/AI models. The models were trained on CONFESS LAI products from the years 2000–2014 as ground truth and land use datasets of LUH2h and HILDA+ as inputs. To ensure the model reflects regional specifications and to avoid a very large model, the globe was divided into smaller regions where models were trained independently. Multiple models were trained, and from different configurations, a variation of XGBoost was chosen as the model structure.
The trained models are used with the input data to infer the CERISE LAI datasets from the year 1925 to 1999, using the land use datasets as inputs to generate the datasets. The CERISE LAI was evaluated against the CONFESS LAI 1-km resolution and the AVHRR-GEO2 4-km resolution for the years 1982–1999. So some key points achieved in this task are as follows:

- An ML/AI framework that successfully emulates monthly LAI values based on annual land use data

- A harmonization and evaluation framework that generates global LAI datasets based on the regional LAI emulator
- A time-varying LAI dataset created from 1925 to 1999 that is compatible with CONFESS LAI and IFS

The results show that model outputs were consistent with the available evaluation and demonstrated better harmonization than the original CONFESS LAI for the years with available data. Despite this, we observed that the results lack interannual variability and converge towards the average in each region, which leads to a reduction of LAI that affects areas with high LAI values (like the Amazon) quite significantly.

“Task 7.4 Create and assess simple representations of time-variation of IFS lakes back to 1925, for reanalysis and possibly for seasonal reforecasts.”

“In CERISE, time-varying lake cover datasets were created based on an open, up to date, consistent in time high horizontal (30 m) and temporal (one month) resolution dataset from Joint Research Centre (JRC) Monthly Water History v1.4 (contains global surface classification maps from 1984 to 2021).”

Generated lake cover maps are global, monthly per decade at 1 km horizontal resolution, and available for 1925-2021:

- 1925-1961 (maps use the same 1962-1971 monthly water distribution, due to lack of information on anthropogenic involvement or other changes in water distribution),
- 1962-1991 (maps use in general 1992-2001 monthly water distribution with regional update based on available reliable satellite information or historic records), and
- 1992-2021 (maps are fully independent and are purely based on satellite information).

Maps have a constant ocean assumption, i.e. all islands built after 1925 (e.g. islands near Dubai, UAE and Singapore) are flooded with lake water till the time they are built when they become land, all coastal changes due to erosion might also become covered with lake water.

Maps were assessed:

- directly by comparison with most reliable available yearly datasets (i.e. global and regional comparison with ESA CCI (300 m), Copernicus CGLS (100 m), ESA WorldCover (10 m)), check of seasonality with dry and rainy season climatology - results show good correlation for the available years,
- indirectly by running offline open-loop (i.e. no data assimilation) IFS experiments (surface module was adopted to use monthly maps instead of single static map), and their results were compared with CCI LAKES daily satellite based skin temperature 1 km resolution product - use of monthly lake covers show improvement over 50 available lakes globally in BIAS and RMSE.

1.2.3 Deviations and countermeasures

No important deviations have been encountered. The deliverable was submitted one month later than planned due to delays in the production of the data for land cover and LAI. Due to time constraints, the land cover classification was evaluated over large regions only. Furthermore, the final CERISE-product is made available at monthly frequency instead of 10-day due to memory constraints in the Machine Learning (ML) algorithm employed for its production. This does not have an impact given that ecLand/IFS only requires monthly data (centered on the middle of month) to do interpolation to daily values.

1.2.4 Reference Documents

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

1.2.5 CERISE Project Partners:

There are 12 project organisation partners active in the CERISE project, which are listed in the following table.

Table 1: *List of the active partners, with the abbreviated and full names, in the CERISE project.*

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

2 Methodology

2.1 Time-varying land cover

2.1.1 Datasets

- **LUH2** (Hurtt et al., 2020) provides detailed information on historical land use and land cover changes. LUH2 includes data on land use transitions and scenarios that are crucial for Earth system models, at yearly timescales (from 850-2100) and 0.25° by 0.25° spatial resolution. In this study, LUH2 is used as one of the primary features to predict land cover changes at a finer resolution. The LUH2 dataset can be found at https://luh.umd.edu/. In this work we use the LUH2h product which covers the 850-2014 period.
- **HILDA+** (Historical Land Use Data) (Winkler et al., 2021), which is a global dataset offering high-resolution (1km) data on historical land use and land cover (1960-2019). HILDA+ captures long-term trends and is used in conjunction with LUH2f to improve the model's ability to predict historical land cover dynamics. The HILDA+ dataset is accessible at https://doi.pangaea.de/10.1594/PANGAEA.921846 .
- **ESA CCI Land Cover** (Lamarche et al., 2017), which provides global land cover classifications at a spatial resolution of 300 meters. The dataset spans the years 1992 to 2020 and includes land cover classes such as forests, grasslands, croplands, and water bodies. The dataset can be accessed at http://maps.elie.ucl.ac.be/CCI/viewer/download.php, and more information is available in the ESA CCI Land Cover Product User Guide at https://www.esa-landcover-cci.org/?q=documents.

2.1.2 Pre-processing

2.1.2.1 Resampling and harmonization of LUH2, ESA-CCI and HILDA+

The Land-Use Harmonization 2 (LUH2) dataset with the native 0.25 deg resolution was directly used for this study without any resampling. The European Space Agency Climate Change Initiative (ESA-CCI) land-cover product, originally available at a native resolution of 300 meters, was resampled to a target resolution of 1 km (~0.0089 degrees) using nearest-neighbor resampling. Following resampling, a mask of the LUH2 grid was generated based on the ESA-CCI grids. This mask allowed for the selection of ESA-CCI grid cells corresponding to each LUH2 cell, thereby enabling direct comparison and integration of the two datasets. To facilitate consistent comparison between LUH2 and ESA-CCI datasets, both were reclassified into a common set of land-use categories. This required the development of a Crosswalk Table (CwT) to translate native land-use classes to a unified classification scheme (Figure 1).

The construction of the harmonized historical land cover maps (From now on, ***CERISE LULC*** involved multiple iterations of permutation and evaluation, where different mappings were tested against the ESA-CCI CwT adapted from Winkler et al. (2020). The final version of the LUH2 CwT was selected based on a quantitative comparison of land-use fractions between LUH2 and ESA-CCI datasets, prioritizing the configuration that produced the best match in land-use distributions across all regions (See Figure 1).

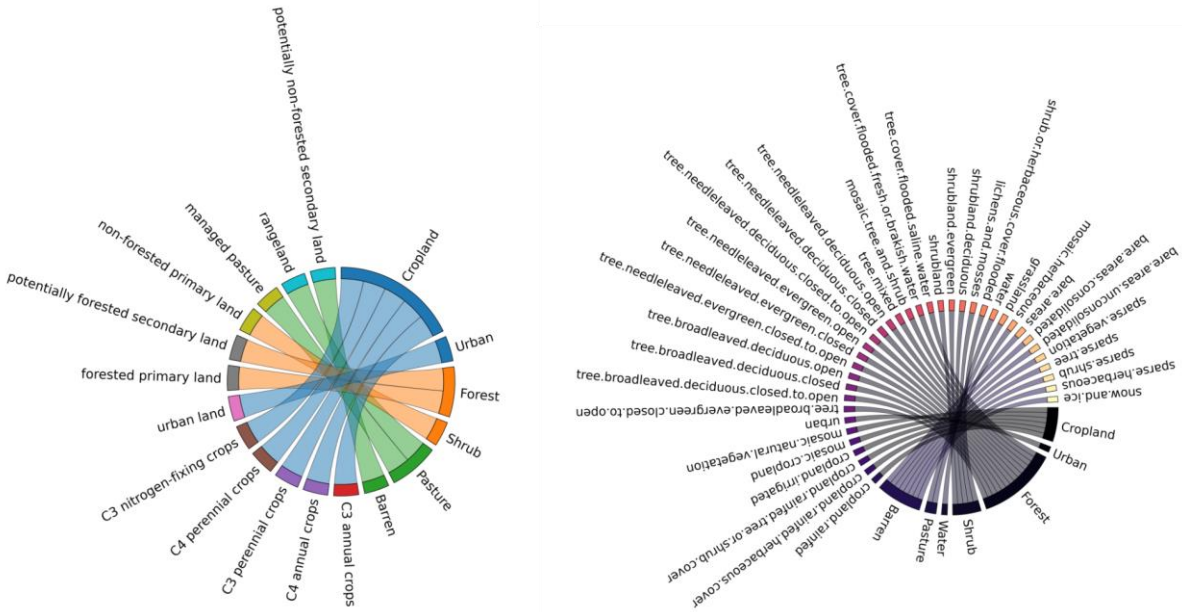

Figure 1: Chord diagrams representing crosswalk table used for reclassifying LUH2 and ESACCI to consistent land use data (This CwT was adapted from Winkler, K et.al. 2020).

2.1.3 Processing

2.1.3.1 LU transitions

After harmonizing the LUH2 and ESA-CCI datasets to a consistent spatial resolution and a unified set of land-use categories, a custom-developed algorithm was employed to simulate annual land-use transitions. This algorithm operates by processing each LUH2 subgrid individually, each of which corresponds to a block of 28 × 28 ESA-CCI grid cells. A schematic representation of the overall methodology is provided in Figure 2.

In this study, the gross land-use transition values provided by the LUH2 dataset were used to simulate land-use change on an annual basis. Prior to the application of transitions, the LUH2 transition files, originally defined according to LUH2-specific land-use categories, were reclassified to match the common land-use types established through the crosswalk table (CwT) described previously. This ensured consistency between the transition information and the reclassified ESA-CCI and LUH2 datasets.

Following the reclassification, a 6 × 6 land use class transition matrix was generated for each LUH2 subgrid and for each simulation year. This matrix captured the magnitude of gross transitions between the six principal land-use classes considered in this study: shrubland, forest, cropland, urban areas, pasture, and barren land. Each element of the transition matrix quantified the area transitioned from one land-use category to another within the LUH2 subgrid.

By default, the spatial location of land use transitions is done by random assignments within each mask, depending on the desired characteristics of the simulation. To ensure that simulated changes in land cover appeared spatially realistic, a spatial prioritization mechanism was added to the algorithm and tested. Specifically, this option identifies boundary cells at the edges of each land-use mask, and preferentially applies changes in land cover at these boundaries. This boundary-focused assignment of land-use transitions resulted in more natural-looking spatial patterns of land-use change in some cases, but in a majority of cases resulted in a grid-like pattern (since the assignment is predominantly done at the boundaries of the LUH2 gridcells), therefore this option was disabled for the final product. Alternatively,

the algorithm was updated to provide an option to perform random assignments within each mask, depending on the desired characteristics of the simulation.

We identified the greatest inconsistencies in pasture-change simulations when the prescribed gross transitions could not be applied due to a dearth of available pasture area - an issue traced to the systematic overestimation of forest and cropland extents in the harmonized LUH2–ESA-CCI-LC baseline. Specifically, ESA-CI-LC has more/less Forest/Pasture compared to LUH2.In such cases, the model initially redirects transitions that were intended to convert pasture into natural land cover (such as forest or shrubland) by instead converting cropland into natural land cover. However, to preserve mass balance and maintain the integrity of transition accounting, these redirected conversions are only carried out after the algorithm first generates or reallocates the necessary amount of pasture area. This approach aims to correct the initial imbalance in transition magnitudes without introducing abrupt land-use changes, while ensuring consistent bookkeeping across all six land-use categories.

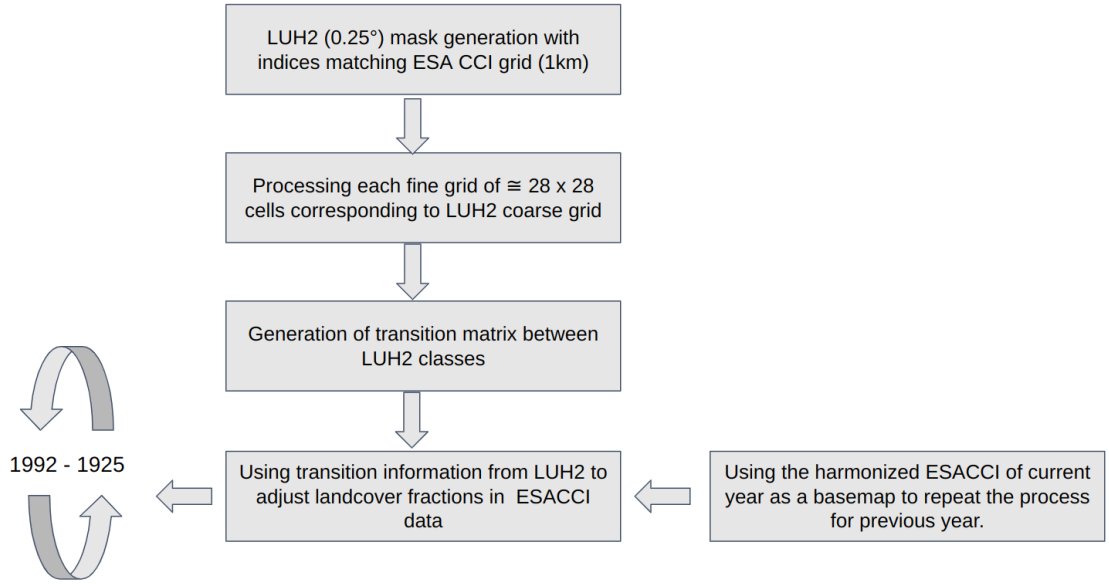

Figure 2: Methodology flow chart illustrating the process flow followed to generate historical land cover maps.

2.1.3.2 LU-LC mapping

The algorithm then applies these land use transition values to the ESA-CCI-derived baseline land cover map. For the initial year, the ESA-CCI land-cover map of 1992 was utilized as the starting basemap. To reconstruct the land-cover state for earlier years (e.g., 1991), the 1992 land-cover map was used in conjunction with the LUH2-specified transitions. Land-use masks were generated for the target year based on the ESA-CCI classification, and within each land-use mask, the most dominant land-cover class from the corresponding 1992 data was identified. Grid cells within each land-use mask that lacked a definitive land-cover assignment were filled by assigning the dominant land cover within that mask. This method was iteratively applied backward in time, generating land-cover maps year-by-year, with each newly reconstructed map serving as the baseline for the previous year.

In certain instances, during the land-use transition application process, the specified transitions from the LUH2 dataset could not be directly mapped to corresponding ESA-CCI grid cells. This typically occurred when the "from-class" defined in the LUH2 gross transition data was absent within the relevant subset of the ESA-CCI land cover grid. To address such inconsistencies, the algorithm incorporated a fallback mechanism designed to ensure continuity in land-cover assignment.

The fallback mechanism relies on a rigorously defined lookup table that links each of the six principal CERISE LULC land-use categories (shrubland, forest, cropland, urban, pasture,

barren) to a single, dominant ESA-CCI land-cover class. These dominant classes were identified by analyzing Table 2 of the ***CONFESS Project Report (2021)***, which quantifies, for each ecLand land-use type, the fractional contributions of all relevant high- and low-vegetation ESA-CCI classes. For each land-use category, we selected the ESA-CCI class with the highest combined vegetation fraction i.e., the class that most closely represents the "average" vegetation structure of that land-use type. During the harmonization process, if the crosswalk table failed to yield a direct one-to-one mapping for a given cell (for instance, due to mismatched or missing class codes), the algorithm automatically assigns this predefined "representative" ESA-CCI class.

The fallback associations between land-use classes and their corresponding dominant land-cover classes are summarized in Table 1 below. In cases where the land-use category did not match any of the predefined classes, a default fallback class of `250` ("Unclassified/Other") was assigned to maintain data integrity.

Table 2: Fallback Mapping Between Land-Use Classes and Land-Cover Classes (Adapted from Table 1 of the CONFESS report).

Land-Use Class ID	Description	Fallback Land-Cover Class ID	Description
1	Forest	30	Tree Cover, Broadleaved, Evergreen
2	Cropland	90	Mosaic Vegetation (Cropland dominant)
3	Urban	190	Artificial Surfaces and Associated Areas
4	Shrubland	30	Shrubland
5	Pasture	110	Tree Cover, Broadleaved, Deciduous, Closed to Open
6	Barren	150	Sparse Vegetation
Other	—	250	Unclassified/Other

2.1.3.3 Map of tiles

To optimize the performance of the land-use harmonization algorithm, the global dataset was partitioned into 32 spatial tiles, enabling distributed and parallel processing on the BSC MareNostrum 5 (MN5) and ECMWF high-performance computing, HPC2020, systems. The number of tiles was empirically determined to balance the computational workload across the available resources, minimizing runtime while ensuring efficient memory usage and

communication overhead. This tiling strategy facilitated scalable execution of the algorithm by allowing each tile to be processed independently (see Figure 3).

On MN5, performance profiling indicated that assigning 110 parallel processes per tile yielded the most efficient execution profile, whereas on ECMWF hpc2020 we used full nodes of 128 parallel processes. This configuration was found to optimally exploit the underlying CPU architecture and memory bandwidth of the HPC nodes. Each tile was submitted as an independent batch job to the SLURM job scheduler, allowing concurrent execution of multiple tiles and thereby significantly reducing the total processing time required for global land-cover simulation.

The algorithm was explicitly developed to leverage MPI-based parallelism via the mpi4py interface, with a focus on minimizing I/O bottlenecks using parallel NetCDF (PnetCDF) I/O operations. Each tile was read, processed, and written independently into tile-specific NetCDF files for each simulation year, ensuring data consistency and parallel I/O efficiency.

Upon completion of tile-wise processing, the output files were systematically mosaicked to reconstruct globally continuous land-cover maps for each year.

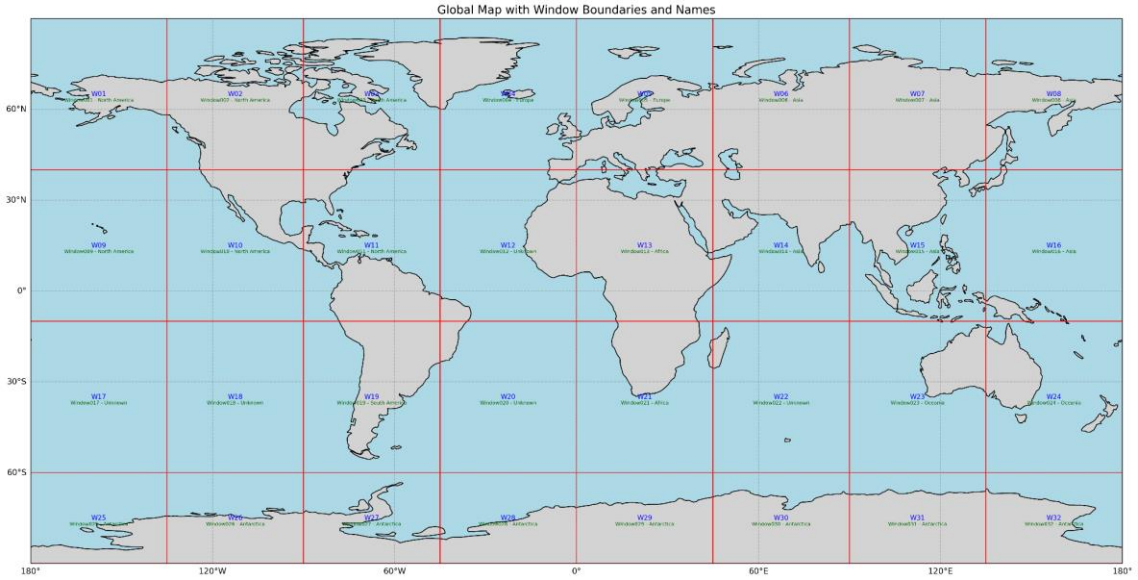

Figure 3: Tiles across the globe used for parallel processing of historical time series landcover.

2.1.4 Post-processing

2.1.4.1 Stitching tiles into global grid

To assemble global land cover maps from regionally processed tiles, the workflow begins by referencing a standard ESA-CCI land cover file to extract the global latitude and longitude grid. A new NetCDF file is initialized using this grid to store the final global outputs across all required years.

Each tile corresponds to a specific region of the globe and contains land cover values stored in a separate NetCDF file, along with a corresponding extent file that indicates the spatial boundaries (row and column indices) of that tile within the global grid. The script reads these extents and places the land cover data from each tile into its correct spatial position within the global array. As a potential improvement, future work could incorporate tile buffers (e.g., a 10-grid overlap) to enhance spatial continuity and reduce boundary artifacts between adjacent tiles.

This process is repeated for all specified tiles, and for each year under consideration. Only the required number of years are copied from the regional files to optimize memory and runtime

efficiency. Once all tiles are integrated, the resulting NetCDF file represents a continuous, spatially complete global land cover map for each year in the simulation period.

2.1.5 Blending lake cover and land cover

The monthly time-varying lake cover data used in Task 7.4 may be inconsistent with the yearly-varying land cover data from Task 7.3, due to different data sources being employed. The ecLand model assumes bare soil if the underlying land cover is a lake or ocean. To avoid any issues when the CERISE land cover assumes there is a lake, whereas there is no lake over the same period in the cover dataset, we apply a corrective measure. The strategy is to identify the points where the situation occurs (which is mainly at the borders of inland water bodies). We then use the 'gdal_fillnodata' tool with the nearest neighbor strategy (available in GDAL version 3.9 and above (Rouault et.al., 2025). In order to avoid propagating to the entire ocean bodies, we use the ESA-CCI permanent land/ocean/inner water maps to mask the ocean points. In Figure 4 we can see the results of filling a number of lake points for the year 2019 of the ESA-CCI derived land cover map around the Northern portion of the Caspian Sea.

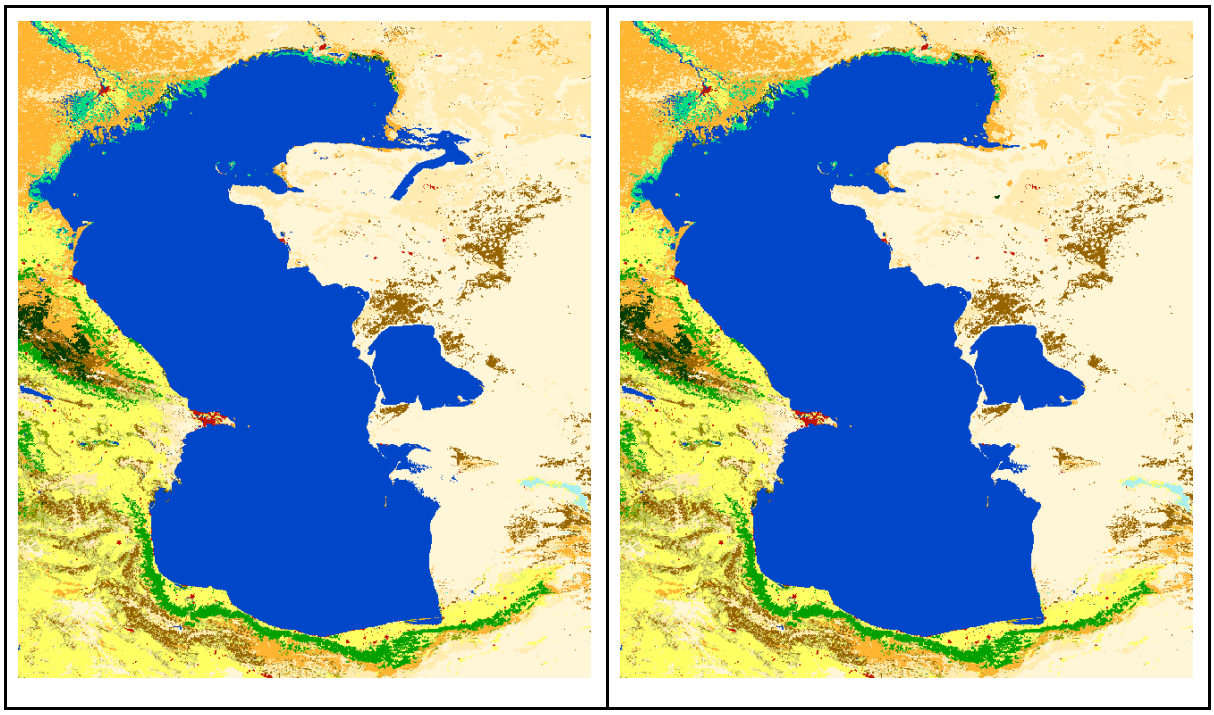

Figure 4: Land cover classification based on ESA-CCI Land Cover in 2019, original data (left) and filled data (right). The Caspian Sea is shown here.

2.1.6 Computational codes

The LC processing code is developed entirely in-house using bash, Python and GDAL utilities, with a focus on high-performance and distributed computing to handle large-scale satellite data. The pipeline is designed to run efficiently on HPC systems by leveraging MPI for parallel execution. It can also be deployed on standard Linux-based clusters or cloud environments using the provided setup files. The codebase utilizes a set of widely adopted scientific and geospatial Python libraries, enabling robust, scalable data manipulation and analysis. The full code and related resources are maintained in the BSC GitLab repository (https://earth.bsc.es/gitlab/es/land-use-reclassification/-/tree/develop).

- mpi4py: A Python wrapper for the MPI standard, enabling distributed parallel processing across multiple compute nodes, critical for handling large Earth observation datasets efficiently.
- xarray: Provides a powerful N-dimensional labeled array structure built on NumPy, tailored for working with large multi-dimensional datasets such as satellite imagery and climate model output.
- netCDF4: A Python interface to the netCDF C library, used for reading, writing, and manipulating netCDF datasets, which are standard formats in Earth sciences.
- NumPy: A foundational package for numerical computation in Python, enabling efficient array operations and data transformations.
- SciPy: Used here mainly for `ndimage.convolve`, supporting multi-dimensional filtering operations needed for spatial data smoothing and processing.
- pickle: Used for serializing Python objects (e.g., models, dictionaries) to disk for reuse across sessions or between processing stages.
- argparse, logging, os, sys, and gc: Standard Python libraries used for argument parsing, logging, system operations, and memory management, respectively.
- GDAL utilities

2.2 Time-varying leaf area index

Machine Learning methods were employed to extend the CONFESS LAI datasets back to 1925 from 1999. The LUH2h and HILDA+ datasets were used as input, and the CONFESS LAI dataset was used as ground truth to train the model.

2.2.1 Datasets

- **LUH2** as mentioned in section 2.1.1
- **HILDA+** as mentioned in section 2.1.1
- **CONFESS LAI** (Boussetta & Balsamo, 2023) products from the CONFESS H2020 project (https://confess-h2020.eu/) that provides global 10-day observation based LAI from 1993 to 2019. Although the project requirement for CONFESS was to generate LAI for the period 1993-2019, an extension back to 1982 was also made available at the end of the project. These LAI products provide detailed and high-resolution data (1km) , which are crucial for accurately modeling vegetation cover and biomass. The use of these products allows us to enhance the model's ability to predict LAI across various spatial and temporal scales.

The CONFESS LAI dataset, used as ground truth for training and predicate for inference, comprises two segments: 1982-1999 and 2000-2014. Due to reliability concerns with the earlier segment over the global tropics identified in CONFESS D1.2 (Improved vegetation variability), only the later years were considered. As LUH2 has multiple segments, the datasets based on observations end in 2014, and for the years after, it uses scenarios. As we didn't want to expose the model to more uncertainties in the training phase, we opted out of using the later years and limited the length of the training to the years (2000-2014) , as the most reliable chunk of CONFESS LAI and LUH2 datasets overlapped.

LUH2 and HILDA+ served as predictors for both training and inference. The earlier years of the CONFESS LAI dataset (1982-1999) were reserved for comparison purposes only.

2.2.2 Pre-processing

- Downscaling of LUH2 Data: The LUH2 dataset was originally provided at a coarser resolution, requiring downscaling to match the grid resolution used in this study. This was achieved by resampling the data to the target resolution using nearest neighbour interpolation through gdalwarp (Rouault et al., 2025). Each subdataset (representing different land-use variables such as cropland, forest, and urban areas) was processed separately, and compression techniques were applied to optimize storage. After processing, all individual variables were combined into a single NetCDF file for each year, ensuring that the downscaled LUH2 data aligned perfectly with the common grid used for the other datasets.
- Remapping of HILDA+ Data: The HILDA+ dataset, originally provided at a 1 km resolution, was remapped to align with the common grid used in this study. Although the resolution remained the same, the remapping ensured that the data was spatially aligned with other datasets. The remapping process was performed using gdalwarp (Rouault et al., 2025) with mode resampling, appropriate for categorical data. Metadata adjustments were made to reflect the changes, and time dimensions were added to enable temporal analysis. This ensured that the HILDA+ dataset was consistent and ready for integration into the final workflow.
- Dynamic water-mask: A yearly binary land-sea mask, generated using HILDA+, serves as input.
- An MLOps pipeline for building a global model by combining independent regional models involves partitioning the globe into regions (It required interpolation due to the different extents of the input and target data.) and creating separate Zarr (Abernathey et al., 2021; Zarr Development Team, 2022) stores for each region to enable efficient and autonomous model training. In Figure 5, global maps including all regions with LAI values are illustrated. Any regions without valid values are discarded and later, in the stitching process, replaced with NaNs. This pipeline facilitates handling large datasets on the accelerated partition without encountering memory constraints, while ensuring that each model captures the unique characteristics of its respective area.

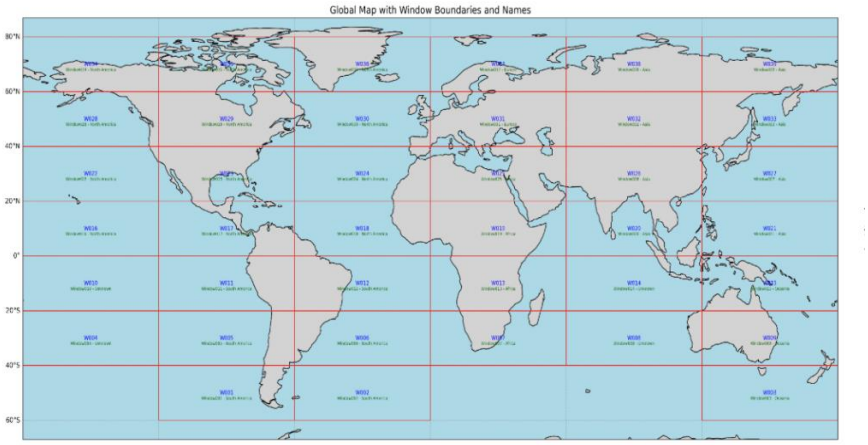

Figure 5: A global model consists of 39 independent regions of 20 °by 60 ° that covers the all the area that CONFESS LAI has any values.

2.2.3 Processing

Three tree-based models have been considered to build the architecture of the processor in CERISE: Random Forest, Extreme Gradient Boosting and Recurrent Neural Network:

- **Random Forest (RF)**, an ensemble learning method from scikit-learn, was introduced by Breiman (2001). It enhances model robustness and accuracy by combining predictions from multiple decision trees. This approach effectively addresses the overfitting issue often encountered with individual decision trees by leveraging diversity in both data and feature selection. RF is a highly effective model and serves as a valuable baseline for comparison with other methods. However, it is not designed for large datasets and the sci-kit learn implementation lacks GPU acceleration support.
- **Extreme Gradient Boosting (XGB)**, created by Chen & Guestrin (2016), is an optimized gradient boosting framework designed to enhance both speed and accuracy in machine learning tasks. It utilizes an ensemble of decision trees, similar to Random Forest, but employs gradient boosting to iteratively minimize prediction errors, making it particularly effective for managing large datasets and capturing intricate data relationships. Furthermore, XGB performs exceptionally well with accelerated GPUs.
- **Recurrent Neural Networks (RNNs)** process sequential data using internal loops and a 'hidden state' to maintain context over time (Elman, 1990). While suitable for tasks like NLP and time series analysis, basic RNNs struggle with vanishing/exploding gradients, hindering their ability to learn long-range dependencies. More advanced variants like Long Short-Term Memory (LSTM) (Hochreiter & Schmidhuber, 1997) and Gated Recurrent Units (GRU) (Cho et al., 2014) employ gating mechanisms to overcome these limitations. In this project, RNN development remained exploratory and did not proceed to a production implementation. Further development and evaluation of these models are planned for future research.

The CONFESS LAI and LULC are used as training datasets spanning the years 2000 to 2014, with LAI as target and LU/LC datasets as inputs. To ensure the model's robustness and accuracy, a temporal split was employed, with the year 2000 designated as the testing period, while the subsequent years, 2001 to 2014, served as the training dataset. The input data incorporated into a single Zarr file that comprised the 14 LUH2h 14 fractional variables, single-value land use data from the HILDA+ dataset, a binary water mask, and importantly, the LAI values from the subsequent year. This input data structure is efficient for the model architecture that we have chosen and is used in an auto-regressive (AR) manner . For instance, when predicting LAI for the year 2000, the LAI from 2001 was used as an input feature. This reflects the model’s backward temporal reconstruction, where each year’s prediction leverages information from the following year. The integration of prior year LAI values was motivated by the aim to enhance the model's temporal consistency and mitigate potential issues such as oscillation and shifting, which could compromise the reliability of the model's predictions over time.

To lessen the extreme skewness of the data (as shown in figure 6), log transformation and normalization to a maximum value of 7 was applied to all LAI data, including auto-regressive LAI. Additionally, a separate region-based water mask was created using HILDA+ water data. Following this, two Zarr files were produced: an input file with 27 stacked feature layers ,

resulting from the concatenation of LUH2 (14 variables), HILDA (1 variable), and auto-regressive LAI (12 variables representing the months), and a ground truth file containing 12 layers (representing the months), one for each year and region.

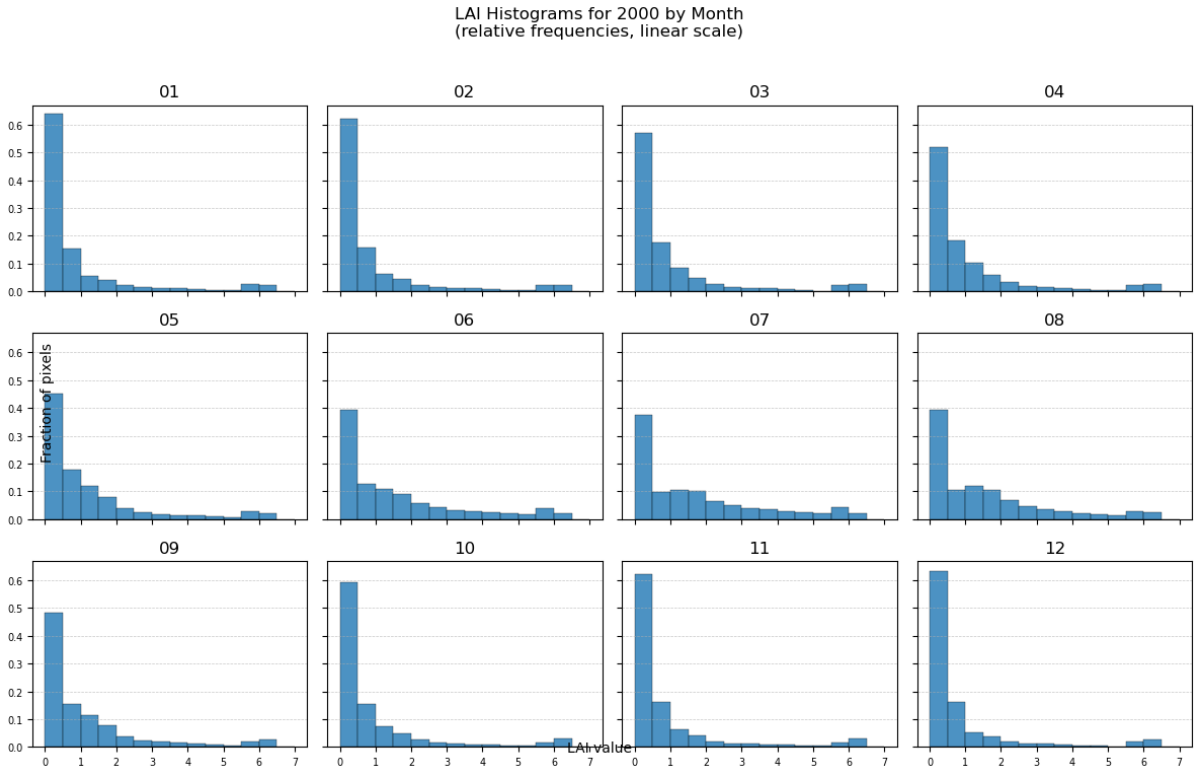

Figure 6: Monthly histograms of LAI for the year 2000 (3×4 grid), binned 0–7 with relative frequencies summing to 1, showing low-LAI distributions in winter, a rightward shift to peak canopy density in summer, and a return to lower values in autumn, averaged across in northern hemisphere.

The model output is transformed back from log scale after the prediction is made and stored as a zarr file.

The tree-based models, RF and XGB, performed very well in our test cases, showing a 91% and 89% reduction in RMS, respectively, across the areas tested. Due to a later start compared to the other two models, the RNN model development could not deliver a viable model by the deadline. Despite showing promising results and utilizing a modern PyTorch implementation that allows for the exploration of more advanced architectures, the RNN model remains experimental and will be further investigated in future work.

The choice between RF and XGB was primarily based on XGB's python implementation (https://xgboost.readthedocs.io/) ability to leverage GPU acceleration and support Dask (https://www.dask.org/), whereas the RF implementation in scikit-learn (https://scikit-learn.org/stable/) does not offer these capabilities. Despite RF showing slightly better performance, its significantly longer training time (approximately six times longer for an identical test case) led us to select XGB as our preferred model. Figure 7 presents the loss function curves for the training and validation sets across XGB boosting rounds, along with the distribution of the training and testing data.

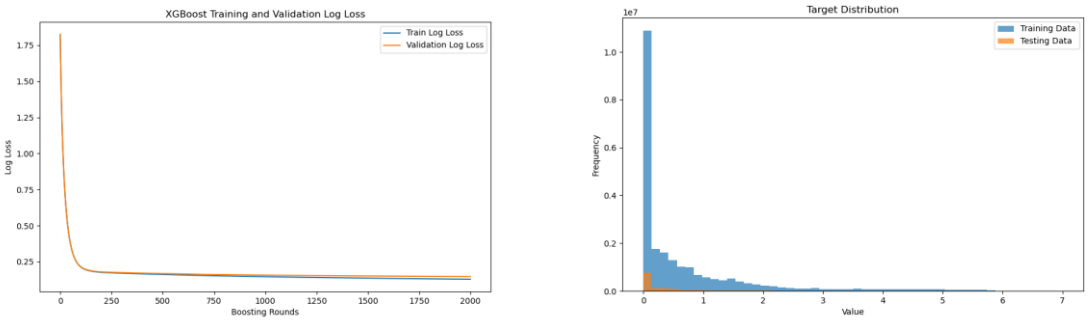

Figure 7: left) The log loss of the XGB for training and validation data after 2000 round of boosting, right) the distribution of the training and validation test data for test studies of over the Iberian Peninsula.

2.2.4 Post - processing

As all the models are trained and inferred independently, a routine was created to collect each year's windows and stitch them together to produce a single prediction per month for each year. To achieve this, we duplicated the original CONFESS LAI NetCDF files, adopted the necessary metadata, and populated the duplicated files with the stitched global data. Additionally, to correct for the land-sea mask, we used the LSM files corresponding to the periods to remove any artifacts in the sea area. Finally, the validation routine was carried out to generate comparisons with the CONFESS LAI at both global and regional scales.

2.2.5 Infrastructure

The computational experiments in this work were performed on MareNostrum 5, a pre-exascale EuroHPC supercomputer at BSC-CNS, with a peak performance of 314 PFlops. The system includes several partitions optimized for different workloads, notably the General-Purpose Partition (GPP) for CPU-based processing and the Accelerated Partition (ACC) for GPU-accelerated tasks. This flexible architecture supports efficient execution of workflows combining data processing with machine learning. The GPP was used for pre-processing and post-processing, taking advantage of its high-core-count CPU nodes, large memory capacity, and fast interconnect. This configuration allowed effective data preparation and feature extraction, ensuring smooth handling of large datasets. The ACC partition, with nodes combining CPUs and Nvidia Hopper GPUs, was employed for training and inference. The GPU acceleration provided the necessary scalability and performance to handle intensive model training and prediction efficiently, reducing computation time while supporting high model complexity.

2.2.6 Code

The code used for the production of the LAI datasets is developed entirely in-house, primarily in Python and Bash, utilizing well-known Python libraries, some of which are listed below. It is designed to run on an HPC system due to the sheer volume of data and the need for GPU access. However, it can be deployed on any Linux machine (cluster or cloud) using the custom Conda environment provided alongside the code. The complete pipeline for both LAI and LC, including all processing steps, is available on the BSC GitHub repository: https://earth.bsc.es/gitlab/ces/ai4land_public.

- zarr**:** A format and library for the storage of chunked, compressed, N-dimensional arrays, designed for scalability in parallel and cloud-based computing environments.
- netCDF**:** A self-describing, machine-independent data format and software libraries that support the creation, access, and sharing of array-oriented scientific data.
- PyTorch**:** An open-source ML library based on the Torch library, providing GPU acceleration and dynamic computation graphs for building and training DL models.
- xgboost**:** An optimized gradient boosting framework that is efficient, flexible, and portable, designed to provide parallel tree boosting for speed and performance.
- scikit-learn**:** An ML library in Python offering efficient tools for classification, regression, clustering, dimensionality reduction, and model evaluation.
- CUDA**:** A parallel computing platform and programming model developed by NVIDIA that enables general-purpose GPU acceleration for compute-intensive applications.

The figure 8 illustrates how training scales with PyTorch on Nvidia H100 GPUs.

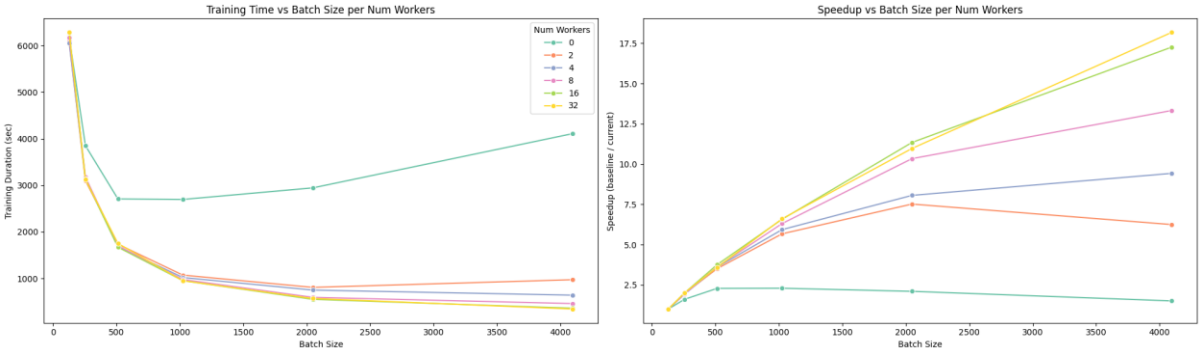

Figure 8: Training time and speedup across varying batch sizes and number of workers for region 25, using 3 epochs over 5 years of data (4 training, 1 validation) on MN5 NVIDIA H100 GPU.

2.3 Time-varying lake cover

To develop a lake cover dataset that capture lakes seasonality and evolution in the current changing climate, two key elements are needed: (i) global reliable vast consistent in time and space high resolution **datasets** based on observations, and (ii) reliable reproducible flexible understandable automated **methodology**, that relies on minimum human intervention, yet very accurate.

2.3.1 Datasets

The main source to generate time-varying lake (i.e. inland water) cover is the 30 meter (1") horizontal resolution (grid EPSG:4326) global (except Antarctica and far North; available 78°N-60°S) water surface dataset from the Joint Research Centre (JRC; Pekel et al., 2016). This dataset was created by using Landsat 5, 7 and 8 individual full-resolution 185 sq.km global reference system II satellite images over the past 38 years to map the spatial and temporal variability of global surface water and its long-term changes. The JRC global water surface dataset consists of several types of maps that show different facets of surface water dynamics on the Earth between the 16th March 1984 and the 31st December 2021. For our purposes, the following maps from the most recent dataset's version 1.4 are the best suited:

- '*Water Transitions*' maps show changes in water classes (no water, seasonal water, permanent water, etc. - 10 water classes in total) on the Earth's surface; used to

determine if water is present over the whole period, appeared only recently, or have any recurring pattern;
- *'Metadata'* maps show number of valid observations and number of times when water was observed; used to specify water class of uncertain geographical locations;
- '*Monthly History*' maps show the entire history of water detection on a month-by-month basis. The collection contains 454 images (i.e. one for each month between March 1984 and December 2021) with water/ notWater/ noData information and used to generate monthly water covers and to further specify water classes based on relevant time periods (e.g. 2012-2021).

Several additional sources were also used to generate the CERISE lake cover:

- Copernicus DEM GLO-30: Global 30m Digital Elevation Model (GLO30, https://doi.org/10.5270/ESA-c5d3d65) has a global (except Armenia and Azerbaijan) 90°N-90°S coverage, at 30 meter horizontal resolution (grid EPSG:4326), and represents year 2015 (uses data for 01.12.2010–31.01.2015). The Copernicus DEM is derived from an edited Digital Surface Model named WorldDEM&trade, i.e. flattening of water bodies and consistent flow of rivers has been included; additional editing of shore- and coastlines, special features (e.g. airports, implausible terrain structures). The product is based on the radar satellite data acquired during the TanDEM-X Mission; used '*Water Body Mask*' that shows 4 surface classes (i.e. not water, ocean, lake, river) to separate ocean from inland water, and *'Elevation'* to update regional water body boundaries 1962-1971;
- numerous regional glacier datasets at 15-100 meter resolution (different grids) representing period 1992-2020, from several sources, i.e. British Antarctic Survey, QUANTARCTICA, GIMP project, QGREENLAND, Norwegian Institute, Icelandic Met service. These datasets were used to improve water distribution over relevant regions;
- Landsat Global Land Survey 1975 (GLS1975, courtesy of the U.S. Geological Survey) has a global (except where data not available) 90°N-90°S coverage, at 60 meter horizontal resolution (grid EPSG:32645). It represents year 1975, but it is based on data for 25.07.1972–20.02.1983. The Global Land Survey (GLS) 1975 is a global collection of imagery from the Landsat Multispectral Scanner (MSS). Most scenes were acquired by Landsat 1-3 in 1972-1983. Data gaps have been filled with scenes acquired by Landsat 4-5 in 1982-1987. These datasets were used to generate regionally monthly water covers 1972-1981.

For the full list of input datasets used to produce the CERISE time-varying lake cover maps please see Table 3.

Table 3: Full list of input datasets used for the CERISE time-varying lake cover map generation.

Dataset [short_name]	***Description (region, resolution, period, format, access)***	***Use & Pre-processing***
Norwegian institute data for Svalbard [**Svalbard**]	Svalbard, 45m resolution, relevant for 2021, in shapefile format. Discrete **glacier** extent data was downloaded from Norwegian institute database - **svalbard_ice**	Correcting water over Svalbard by removing wet glacier tops with svalbard_ice during Water_History data use.
Icelandic Meteorological Office data for Iceland [**Iceland**]	Iceland, 100m resolution, relevant for 2017, in GeoTiff format. Fractional **glacier** cover was required from personal communication with Bolli Palmason in 2019 - **iceland_ice**	Correcting water over Iceland by removing wet glacier tops with iceland_ice during Water_History data use.

QGreenland [**QGRL**]	Greenland, 30m resolution, relevant for 2017, in GeoTiff format. So far is the most reliable source of **lake** location for the region (in total information for 4528 lakes); was downloaded from QGreenland website - **qgrl_lake**	Rasterizing Greenland lakes qgrl_lake for water correction.
The Greenland Ice Mapping Project data [**GIMP**]	Greenland, 15m resolution, relevant for 2002 [1999.06.30-2002.09.04], in GeoTiff format. **Ice** cover mask is a satellite composite product using Landsat 7 ETM+ imagery and RADARSAT-1 SAR amplitude images; was downloaded from National Snow and Ice Data Center - **gimp_ice** (0 - not glacier ice, 1 - glacier ice).	Correcting water over Greenland by filling all inland water with land, and putting on top only proglacial lakes (i.e. lakes from ggrl_lake that do not intersect with gimp_ice) during Water_History data use.
QuAntarctica [**QANT**]	Antarctica, 30m resolution, relevant for 2021, in shapefile format. High quality (higher than BAS) source on Antarctic (and additionally more detailed for the East Antarctica and Schirmacher oasis) proglacial **lakes** location and area (ignored lakes with wrong location and area less than 0.01 sq.km; downloaded from QuAntarctica website - **qant_lake**, **qeast_lake**	Rasterizing Antarctica proglacial lakes qant_lake and qeast_lake for water correction.
British Antarctic Survey v7.6 & v7.3 [**BAS**]	Polar Antarctic zone (60°S-90°S), South Georgia and the South Sandwich Islands (50°S-60°S), 2m & 30m resolution, relevant for 2022 [2022.11.11–2022.11.11], in shapefile format. Antarctic Digital Database (ADD) compiles the best available geographic information covering south of 60°S. Following datasets are used: (i) Medium resolution vector polygons of the Antarctic coastline v7.6 (2022.11.11) - **coastline** for Antarctica ('land', 'ice shelf', 'ice tongue' or 'rumple' attribute), used for surface detection by type and location (i.e. islands); all surface types considered as land and fully covered with **glacier**; The South Georgia GIS is a collection of topographic, management and scientific datasets about South Georgia and the South Sandwich Islands, used for surface detection by type and location (i.e. islands) and correction of artificial islands 50°S-60°S (fields 'seamask' - polygon excludes land (& islands), 'coastline' - polygon includes land (& islands)); downloaded from BAS website - **Antarctica_coast**, **AntRegion_coast** (coastlines for all islands of the South Georgia region), **AntRegion_ice** ('Ice', 'Lake', 'Moraine', 'Ice-free'), **AntRegion_lake** ('Ice', 'Lake', 'Moraine', 'Ice-free'; 444 lakes), **AntRegion1_lake** (lake detailed information for small Barff region), **AntRegion2_lake** (Busen region), **AntRegion3_lake** (Thatcher region)	Rasterizing the South Georgia and the South Sandwich Islands coastline AntRegion_coast - to correct MERIT DEM elevation data merit30 over that region. Merging and rasterizing (i) Antarctica and the South Georgia and the South Sandwich Islands coastlines (assumed as fully land), i.e. Antarctica_coast and AntRegion_coast - to correct pixel type for land over the region during Water_History data use; (ii) Antarctica coastline (assumed as fully covered with glaciers) and the South Georgia and the South Sandwich Islands ice cover, i.e. Antarctica_coast and AntRegion_ice - to correct pixel type for land over the region during Water_History data use; (iii) Antarctica and the South Georgia and the South Sandwich Islands lake cover, i.e. qant_lake, qeast_lake, and AntRegion_lake, AntRegion1_lake, AntRegion2_lake, AntRegion3_lake - to correct pixel type for water over the region during Water_History data use.
Global Land Ice Measurements from Space project data [**GLIMS**]	Global (except where data not monitored), 30m resolution, relevant for 2023 [1750.01.01-2023.06.07], in shapefile + table format.	Merging rasterized ice covers for Svalbard svalbard_ice, Iceland iceland_ice, Greenland gimp_ice, Antarctica Antarctica_coast, the South

	Global **ice** cover comprises of 1186805 total records (730338 not 'gone'; downloaded from Google Earth Engine (GEE) catalogue - **glims_ice**	Georgia and the South Sandwich Islands AntRegion_ice - to correct pixel type for land over the region during Water_History data use.
MERIT DEM: Multi-Error-Removed Improved-Terrain DEM [**MERIT DEM**]	Global (except Antarctica), 90°N-60°S, 90m resolution, relevant for 2017 [1987.01.01–2017.01.01], in GeoTiff format. MERIT DEM is a high accuracy global DEM at 3 arc second (~90 m at the equator) resolution produced by eliminating major error components from existing DEMs (NASA SRTM3 DEM, JAXA AW3D DEM, Viewfinder Panoramas DEM). Significant improvements were found in flat regions where height errors were larger than **topography** variability, and landscapes such as river networks and hill-valley structures became clearly represented; downloaded from Yamazaki Lab MERIT DEM website - **merit** (elevation data in meters)	Correcting wrongly allocated islands near South Georgia and the South Sandwich Islands (East of the southern point of South America) with the rasterized coastline from BAS data, i.e. Antarctica_coast and AntRegion_coast (islands with no elevation data are marked as -9999) to generate merit30.
JRC Monthly Water History [**Water_History**]	Global (except Antarctica, far North), 78°N-60°S, 30m resolution, monthly, relevant for 1984-2021 [1984.03.16–2022.01.01], in GeoTiff format. This dataset contains maps of the location and temporal distribution of **surface water** from 1984 to 2021 and provides statistics on the extent and change of those water surfaces. Monthly History collection holds the entire history of water detection on a month-by-month basis. Areas where water has never been detected are masked; download from GEE catalogue - **water_history** (0: No data, 1: Not water, 2: Water; 454 maps), **transition** (categorical classification of water change between first and last year of available data; 0:NoChange, 1:Permanent, 2:NewPermanent, 3:LostPermanent, 7:SeasonalToPermanent)	Creating permanent water map based on transition layer by selecting only 1:Permanent, 2:NewPermanent, 7:SeasonalToPermanent water types - to fill the missing data over Armenia and Azerbaijan in the water body mask glo30_wbm. Calculating water_class map based on monthly Water_History data for the past 10 years (2012-2021) by counting how many times each pixel was marked 0:NoData, 1:NotWater, 2:Water (in total each type could be max 120 times, in reality due to missing data much less), then marking pixels with greater number of 1:NotWater as land, 2:Water as water, and rest as 120 (only 0:NoData or number of land and water were equal) - to correct the gap-filled water body mask glo30_wbm.
Landsat Global Land Survey 1975 [**GLS1975**]	Global (except where data not available), 60m resolution, relevant for 1975 [1972.07.25-1983.02.20], in GeoTiff format. The Global **Land Survey** (GLS) 1975 is a global collection of imagery from the Landsat Multispectral Scanner (MSS). Most scenes were acquired by Landsat 1-3 in 1972-1983 (data gaps have been filled with scenes acquired by Landsat 4-5 in 1982-1987). These data contain 4 spectral bands: Green, Red, an NIR band, and a SWIR band; downloaded from GEE catalogue - **gls1975** (short-wavelength infrared (800-1100 nm))	Creating permanent water map for the 1972-1981 period based on gls1975 by selecting only short-wavelength infrared layer values less than or equal 12.
Copernicus DEM GLO-30: Global 30m Digital Elevation Model [**GLO30**]	Global (except Armenia and Azerbaijan), 90°N-90°S, 30m resolution, relevant for 2015 [2010.12.01–2015.01.31], in GeoTiff format. The Copernicus DEM is a Digital Surface Model (DSM) representing the Earth's surface, including buildings, infrastructure and vegetation. Derived from an edited DSM named WorldDEM&trade, i.e. flattening of	Filling the missing data of the water body mask glo30_wbm: (i) over Armenia and Azerbaijan with permanent water based on transition layer; (ii) over 89°-90°S with only land. Correcting the gap-filled water body mask glo30_wbm with (i) inland_water over Maracaibo Lake and Azov Sea; (ii)

	water bodies and consistent flow of rivers has been included. The WorldDEM product is based on the radar satellite data acquired during the TanDEM-X Mission; downloaded from GEE catalogue - **glo30_wbm** (water body mask (0=NoWater, 1=Ocean, 2=Lake, 3=River) with missing Armenia, Azerbaijan), **glo30_elv** (elevation data in meters with missing Armenia, Azerbaijan), **glo90_elv_region** (elevation data from Copernicus catalogue over Armenia, Azerbaijan at 90m)	land over water_class(land) and merit30(elevation greater 0m or equal -9999).
Large Scale International Boundary Polygons, Detailed [**LSIB**]	Global, 90°N-90°S, 30m resolution, relevant for 2017 [2017.12.29-2017.12.29], in shapefile format. LSIB is derived from (i) LSIB line vector file, and (ii) World Vector Shorelines (WVS) from the National Geospatial-Intelligence Agency (NGA). The interior boundaries reflect U.S. government policies on boundaries, boundary disputes, and sovereignty; downloaded from GEE catalogue - **LSIB** (attributes used are 'COUNTRY_NA' US-recognized country name, 'OBJECTID', 'Shape_Area'; in total 284 countries/ 180,741 features)	Calculating missing island mask by masking pixels of the corrected water body mask glo30_wbm which belong to any LSIB country.

2.3.2 Processing

The main idea of the developed methodology is to use only most necessary input datasets and be easily adaptable if the new better dataset or dataset's version comes out.

In short, time-varying lake covers were generated in the following way. First, we determine the dominant grid cell type (i.e. water, notWater, noData) for (i) the whole period, (ii) each month of the whole period, (iii) every 10 years of the whole period. Second, we fill the noData grid cell type by combining previously obtained information. Third, we calculate permanent water distribution per 10-year period and (i) correct regionally in space over glaciers, islands, and far north areas, (ii) correct regionally in time for years prior to available data. Then, we calculate seasonal monthly water distribution per 10-year period following the same procedure as for permanent water. Finally, we separate water into inland and ocean, and average data from 30 meters to 1 km.

For all the details on methodology please see Table 4, and for details on exceptional areas please see Table 5.

Table 4: Detailed description of time-varying lake methodology.

***Dataset used* / Step description / Step output**
Monthly History Calculating number of (1: Not water) and (2: Water) occurrences per pixel over 1984-2021 (i) all data, (ii) each month data. Output: number of 'NotWater' 13 global maps, 30m resolution regular grid; number of 'Water' 13 global maps, 30m resolution regular grid

Monthly History

Determining the type of each pixel over 1984-2021 (i) all data, (ii) each month data: (1) definitely (1: Not water) - number of (1: Not water) is greater than number of (2: Water), and number of (2: Water) is less or equal 10 and less or equal 5% of total (1: Not water) and (2: Water) sum; (2) definitely (2: Water) - number of (2: Water) is greater than number of (1: Not water), and number of (1: Not water) is less or equal 10 and less or equal 5% of total (1: Not water) and (2: Water) sum; (3) surface type - '0' definitely (1: Not water), '1' definitely (2: Water), '-1' where criteria were not met (undecided).

Output: surface type (i.e. 0 'NotWater', 1 'Water', -1 'Undecided') 13 global maps, 30m resolution regular grid

Monthly History

Updating the type of each pixel over 1984-2021 monthly data - filling '-1' (undecided) values of all data with appropriate monthly data.

Output: surface type (i.e. 0 'NotWater', 1 'Water', -1 'Undecided') 12 global maps, 30m resolution regular grid

Monthly History

Determining the type of each pixel over each 10-year period of 1982-2021:
(1) calculating number of (1: Not water) and (2: Water) occurrences per pixel over each month data;
(2) calculating of (1: Not water) and (2: Water) occurrence percentages of total (1: Not water) and (2: Water) sum per pixel over each month data;
(3) updating (2: Water) occurrence percentage with water based on 1984-2021 appropriate month values;
(4) updating (1: Not water) occurrence percentage with land where (2: Water) occurrence percentage equals 100%;
(5) determining water presence per pixel over each month data - '1' if (2: Water) occurrence percentage is greater or equal 75%, '0' otherwise;
(6) determining permanent water presence per pixel - minimum water presence within 12 months, filled with water where (1: Not water) occurrence percentage is less then 10%.

Output: surface type (i.e. 0 'NotWater', 1 'Water', -1 'Undecided') 4x13 global maps, 30m resolution regular grid

Monthly History, regional glacier datasets (Svalbard, Iceland, QGRL, GIMP, QANT, BAS, GLIMS), Large Scale International Boundary Polygons, Detailed [LSIB], Copernicus DEM GLO-30: Global 30m Digital Elevation Model [GLO30]

Updating permanent water presence (i.e. type of each pixel on the permanent water map) over each 10-year period of 1982-2021:
(1) filling with land over Antarctica and the South Georgia and the South Sandwich Islands (over Antarctica_coast and AntRegion_coast), over Greenland, and over global ice cover (over Svalbard svalbard_ice, Iceland iceland_ice, Greenland gimp_ice, Antarctica Antarctica_coast, the South Georgia and the South Sandwich Islands AntRegion_ice);
(2) filling with water over Antarctica and the South Georgia and the South Sandwich Islands (i.e. gant_lake, geast_lake, and AntRegion_lake, AntRegion1_lake, AntRegion2_lake, AntRegion3_lake), over Greenland (i.e. ggrl_lake);
(3) (extra only for 10-year periods of 1982-2011) due to less valid observations in earlier periods some islands might become missing - fill missing islands with 2012-2021 values (i.e. missing islands are areas with land/lake/river over corrected water body mask glo30_wbm that do not belong to any country based on country mask lsib).

Output: surface type (i.e. 0 'NotWater', 1 'Water', -1 'Undecided') 4(periods)x1 global maps, 30m resolution regular grid

Monthly History, GLO30

Updating permanent water presence (i.e. type of each pixel on the permanent water map) over each 10-year period of 1982-2021:
(1) due to low number of valid observations and negligible water variation in time some areas at the far north and island Sao Tome (near Africa) were replaced with corrected water body mask glo30_wbm;
(2) (extra only for 10-year periods of 1982-2011) due to less valid observations in earlier periods and negligible water variation in time some regions (i.e. Iceland, Faroe Islands, Novaja Zemla island, Taymyrsky Kraj region, Anju Islands, Wrangler island, and South-West Africa) were replaced with 2012-2021 values;
(3) (extra only for 10-year periods of 1982-2001) due to less valid observations in earlier periods and strong anthropogenic impact some regions (i.e. islands near Dubai, United Arab Emirates) were replaced with water;
(4) (extra only for 10-year period of 1982-1991) due to negligible number of valid observations in earlier period only some regions were updated with 1982-1991 data (i.e. Great Salt Lake in USA, Lake Mead in USA, Lake Poopo in Bolivia, Lake Chad in Chad/Cameroon/Nigeria/Niger, Dead Sea in Israel/Jordan, Lake Habbaniyah in Iraq, Lake Milh in Iraq, Lake Urmia in Iran, Aral Sea in Kazakhstan/Uzbekistan, Menindee Lakes in Australia; and large regions surrounding Great Salt Lake in USA, Mississippi River in USA, Brazil, Argentina, Australia), rest remained the same as 1992-2001 period;
(5) separating water into inland and ocean (i.e. based on the corrected water body mask glo30_wbm globally, except Antarctica - based on overlap and polygon in/out detecting algorithms);
(6) reducing resolution with mean to a 1-km regular grid (EPSG:4326) resolution.

Output: water cover 2(all water, only inland water)x4(periods)x1 global maps, 30m resolution regular grid; water cover fraction 2(all water, only inland water)x4(periods)x1 global maps, 1km resolution regular grid

Landsat Global Land Survey 1975 [GLS1975], GLO30

Updating permanent water presence (i.e. type of each pixel on the permanent water map) over each 10-year period of 1962-1981:
(1) (extra only for 10-year period of 1972-1981) due to high noise level of the short-wavelength infrared data from gls1975 only data for certain regions were processed (only values less than or equal to 12 were used), modified (individual for each region) and used (i.e. Great Salt Lake in USA, Lake Mead in USA, Lake Poopo in Bolivia, Lake Chad in Chad/Cameroon/Nigeria/Niger, Dead Sea in Israel/Jordan, Lake Habbaniyah in Iraq, Lake Milh in Iraq, Lake Urmia in Iran, Aral Sea in Kazakhstan/Uzbekistan, Menindee Lakes in Australia), rest remained the same as 1982-1991 period;
(2) (extra only for 10-year period of 1962-1971) due to fragmented and approximate information available (i.e. text description, smoothed printed maps, very little details) only for certain regions (i.e. Great Salt Lake in USA, Lake Mead in USA, Lake Poopo in Bolivia, Lake Chad in Chad/Cameroon/Nigeria/Niger, Dead Sea in Israel/Jordan, Lake Habbaniyah in Iraq, Lake Milh in Iraq, Lake Urmia in Iran, Aral Sea in Kazakhstan/Uzbekistan, Menindee Lakes in Australia) gap-filled elevation data from glo30 is processed, modified (individual for each region) and used, rest remained the same as 1982-1991 period;
(3) separating water into inland and ocean (i.e. based on the corrected water body mask glo30_wbm globally, except Antarctica - based on overlap and polygon in/out detecting algorithms);
(4) reducing resolution with mean to a 1-km regular grid (EPSG:4326) resolution.

Output: water cover 2(all water, only inland water)x2(periods)x1 global maps, 30m resolution regular grid; water cover fraction 2(all water, only inland water)x2(periods)x1 global maps, 1km resolution regular grid

Monthly History Determining seasonal water presence (12 months) over each 10-year period of 1962-2021: NOTE: for periods 1962-1971, 1972-1981, 1982-1991 monthly data for 1991-2001 is used (0) maximum of updated permanent water presence and month's water presence, then following the same steps as for the permanent water presence update; i.e.: (1) filling with land over Antarctica and the South Georgia and the South Sandwich Islands (over Antarctica_coast and AntRegion_coast), over Greenland, and over global ice cover (over Svalbard svalbard_ice, Iceland iceland_ice, Greenland gimp_ice, Antarctica Antarctica_coast, the South Georgia and the South Sandwich Islands AntRegion_ice); (2) filling with water over Antarctica and the South Georgia and the South Sandwich Islands (i.e. gant_lake, geast_lake, and AntRegion_lake, AntRegion1_lake, AntRegion2_lake, AntRegion3_lake), over Greenland (i.e. ggrl_lake); (3) filling missing islands with 2012-2021 values; (4) replacing the far north and island Sao Tome (near Africa) with corrected water body mask glo30_wbm; (5) replacing certain regions (i.e. Iceland, Faroe Islands, Novaja Zemla island, Taymyrsky Kraj region, Anju Islands, Wrangler island, and South-West Africa) with 2012-2021 values; (6) additional regional correction in time for (i) Toshka Lakes in Egypt (were formed only in 1998) - area is filled with land for periods 1962-1971, 1972-1981, 1982-1991, 1992-2001; (ii) islands near Dubai in United Arab Emirates (were built only in 2001-2003) - area is filled with water for periods 1962-1971, 1972-1981, 1982-1991, 1992-2001; (iii) Great Salt Lake in USA (1962-1971 distribution mimics 2012-2021) - for period 1962-1971 area is filled with maximum water distribution between perm_1962-1971 and monthN_2012-2021; (7) separating water into inland and ocean (i.e. based on the corrected water body mask glo30_wbm globally, except Antarctica - based on overlap and polygon in/out detecting algorithms); (8) reducing resolution with mean to a 1-km regular grid (EPSG:4326) resolution. Output: water cover 2(all water, only inland water)x6(periods)x12 global maps, 30m resolution regular grid; water cover fraction 2(all water, only inland water)x6(periods)x12 global maps, 1km resolution regular grid

Table 5: Detailed description of areas with additional correction.

***Location* / Correction source / Correction description**
Great Salt Lake Corrections based on https://pubs.usgs.gov/wsp/2332/report.pdf and details from https://en.wikipedia.org/wiki/Great_Salt_Lake The area of the lake can fluctuate substantially due to its low average depth of 4.9 m. In the 1980s, it reached a historic high of 8,500 km2. In 2021, it fell to its lowest recorded area at 2,500 km2, falling below the previous low set in 1963, due to years of sustained drought and increased water diversion upstream of the lake. The water level of the lake has a yearly cycle - the rise between September and December and the decline between March and July. [https://pubs.usgs.gov/wsp/2332/report.pdf].
Lake Mead Corrections based on https://earthobservatory.nasa.gov/images/45945/water-level-changes-in-lake-mead Reservoir was build in 1930s (still filling in 1937). In August 2010, Lake Mead reached its lowest level since 1956. According to records, the lake held roughly 27.8 million acre-feet of water at its high point in 1941, and levels have fluctuated through drought in the 1950s and the filling of another upstream reservoir, Lake Powell, in the 1960s. Lake levels rose steadily through the 1980s, reaching 24.8 million acre-feet in August 1985. But as of August 2010, Lake Mead held 10.35 million acre-feet, just 37 percent of the lake's capacity.

Lake Poopo

Corrections based on https://www.tandfonline.com/doi/pdf/10.1623/hysj.51.1.98 (Fig.4) and details from https://en.wikipedia.org/wiki/Lake_Poop%C3%B3

The lake lacked any major outlet and has a mean depth of less than 3 m, the surface area differed greatly seasonally.
In 2002, the lake was designated as a site for conservation under the Ramsar Convention. By December 2015, the lake had completely dried up, leaving only a few marshy areas. Despite the lake rebounding from two previous recorded drying instances, as of 2016, the lake's recovery is considered unlikely. Suggested causes of the decline are the melting of the Andes glaciers and loss of their waters, because of a drought due to climate change, as well as continued diversion of water for mining and agriculture. At its maximum in 1986, the lake had an area of 3,500 km2. During the years that followed, the surface area steadily decreased until 1994, when the lake disappeared completely (20.01.2016 - declared a disaster zone). The time period between 1975 and 1992 was the longest period in recent times when the lake had a continuous water body.
Max rainfall in winter, min precipitation (draught) in May-August.

Lake Chad

Corrections based on https://www.grida.no/resources/5593 and details from https://www.britannica.com/list/7-lakes-that-are-drying-up

Historically, surface area varies greatly by season as well as from year to year. When the surface of the lake is ~280 m above sea level, the area is about 17,800 square km. In the early 21st century, the area was typically about 1,500 square km.
The surface area typically reaches its maximum in late October or early November before shrinking by more than half by late April or early May.
The volume of the lake reflects local precipitation and the discharge of its catchment area, balanced against losses through evaporation, transpiration, and seepage.
Since the 1960s, however, Lake Chad has shrunk approximately 90 percent, because of variations in climate and water withdrawals from irrigation.

Dead Sea

Corrections based on images from https://link.springer.com/article/10.1007/s42452-020-2146-0 and details from https://en.wikipedia.org/wiki/Dead_Sea

Since 1930, when its surface was 1,050 km2 and its level was 390 m below sea level, the Dead Sea has been monitored continuously.
The Dead Sea has been rapidly shrinking since the 1960s because of diversion of incoming water from the Jordan River to the north as part of the National Water Carrier scheme, completed in 1964. As of 2021, the surface of the Sea has shrunk by about 33 percent since the 1960s.

Lake Habbaniyah and Lake Milh

Information for corrections is based on https://www.researchgate.net/publication/329983297_Comparison_of_derived_Indices_and_unsupervised_classification_for_AL-Razaza_Lake_dehydration_extent_using_multi-temporal_satellite_data_and_remote_sensing_analysis and details from https://earthobservatory.nasa.gov/images/147315/iraq-lakes-bounce-back

Lake Milh was constructed in 1970s, but there is no indication what was there before that. NOTE: 1962-1971 period is identical to 1972-1981 as no better information is available.

Lake Urmia

Corrections based on images from https://www.theguardian.com/world/iran-blog/2015/jan/23/iran-lake-urmia-drying-up-new-research-scientists-urge-action and detailed analysis from https://www.bbc.com/future/article/20210225-lake-urmia-the-resurrection-of-irans-most-famous-salt-lake

Following the 1979 revolution, which overthrew the monarchy, Iran adopted a policy of food self-sufficiency and started growing irrigation-intensive crops.

For a while, until about 1995, the lake appeared to be just about holding on its own despite low rainfall since the 1970s. Things began to deteriorate quite quickly from there.

Aral Sea

Corrections based on https://www.britannica.com/place/Uzbekistan and extra info https://www.grida.no/resources/5615

Aral Sea is an endorheic lake that used to be one of four largest lakes in the world with water surface area of 68'000 sq.km. Historical records show that shrinking of the Aral Sea started at least in the middle of the 18th century and was accelerated in 1960's after massive diversion of water for cotton and rice cultivation. Shrinking continued. Due to major Aral Sea recovery program launched in 2001 by Kazakhstan President Nursultan Nazarbayev and supported by the World Bank, Aral Sea water surface area in 2008 became more or less stable ~ 3'300 sq.km. [The Kazakh Miracle: Recovery of the North Aral Sea, Environment News Service (ENS) 2008; http://www.ens-newswire.com/ens/aug2008/2008-08-01-01.asp]
NOTE: 1962-1971 water distribution should be constant going backwards in time.

Menindee Lakes

Corrections are based on https://earthobservatory.nasa.gov/images/148336/menindee-lakes-finally-refilling and details from https://www.mdba.gov.au/water-management/infrastructure/menindee-lakes

Work to use the Menindee Lakes for water conservation started in 1949, with major works finished in 1960 and overall completion in 1968.
NOTE: 1962-1971 period is identical to 1972-1981 as no better information is available.

Toshka Lakes

Corrections are based on satellite images and event reports.

Lake only formed in 1998 due to heavy flash floods. Started drying straight away, almost fully dried in 2018. From the middle of 2018 due to constant flash floods and heavy precipitation in Sudan and South Sudan fully regenerated and in August 2023 even formed a new lake.

Dubai Islands

Corrections are based on satellite images and news articles.

Islands were built in 2001-2003.

3 Results

3.1 Land Cover

3.1.1 Timeseries of land use states and transition rates

Figure 9 shows the global-scale comparison of absolute area per primary land-use class produced by our algorithm against LUH2 and HILDA+ historical reconstructions. Because our simulation is initialized in 1992 using the ESA-CCI climatology, any mismatch between modeled and reference extents at this start date is propagated (and typically magnified) when the time series is extended backward. As a result, our model overestimates cropland and forest areas relative to both LUH2 and HILDA+ in the early period, although it nevertheless reproduces the broad temporal trajectories specified by LUH2.

Figure 10 presents the same evaluation restricted to the South American domain—a region of particular interest given the extensive deforestation of the Amazon. Here again we observe systematic overestimation of cropland and forest cover in our simulations compared to LUH2. Despite these amplitude biases, the algorithm successfully captures the temporal trend in land-use change prescribed by LUH2, indicating that the cellular transition rules and spatial allocation procedures are robust even in a region undergoing rapid land-cover dynamics. A notable source of this divergence lies in the harmonization process. During the alignment of LUH2 with ESACCI post-1992 data, no corrective scaling was applied to reconcile the mismatch in absolute land-use fractions at the starting point (1992). This decision was made deliberately to preserve continuity in land-use time series and avoid artificial discontinuities that could arise from abrupt magnitude corrections. As a result, differences in initial land-cover extents, particularly for certain classes, persisted throughout the pre-1992 period.

Despite these initial discrepancies, our algorithm successfully captures the broader temporal trends and climatology associated with each land-use class as represented in LUH2. This is especially evident for the *Forest* and *Crop* categories, where the modeled trajectories closely follow LUH2 in both magnitude and temporal evolution, indicating effective adaptation and transition modeling for these classes.

In contrast, performance varied significantly for other land-use classes—specifically *Shrub*, *Pasture*, and *Urban*. These classes exhibited pronounced deviations from LUH2, beginning with lower initial area magnitudes in 1992. In the case of *Shrub* and *Pasture*, these discrepancies compounded over time, leading to a rapid decline in their respective extents when going back in time. Eventually, these categories disappeared entirely from many subgrids when going back in time, eliminating their availability for further transitions within the modeled system. We attribute this behavior primarily to two factors:

1. **Initial Magnitude Mismatch**: The underestimation of baseline coverage for these classes in 1992 reduced their persistence in the backward-projected time series.

2. **Saturation and Spatial Competition**: Limited spatial availability within subgrids may have favored transitions toward dominant classes (e.g., Forest or Cropland), effectively "saturating" the grid and crowding out minor classes such as Shrub and Pasture. This was partially remedied as explained in the Methodology section, but was not completely effective.

These effects were most pronounced for *Shrub*, *Pasture*, and *Urban* categories, which showed a declining presence and less fidelity to LUH2 trends for past periods. In particular, the absence of Pasture land in subgrids past a certain historical threshold suggests the algorithm's

sensitivity to initial conditions and competitive dynamics among land-use types in constrained spatial environments.

In summary, while the algorithm demonstrates strong capability in aligning with reference datasets for major land-cover types (notably Forest and Cropland), the representation of minor or transitionally unstable classes highlights challenges related to initialization, subgrid spatial dynamics, and the legacy of dataset harmonization choices.

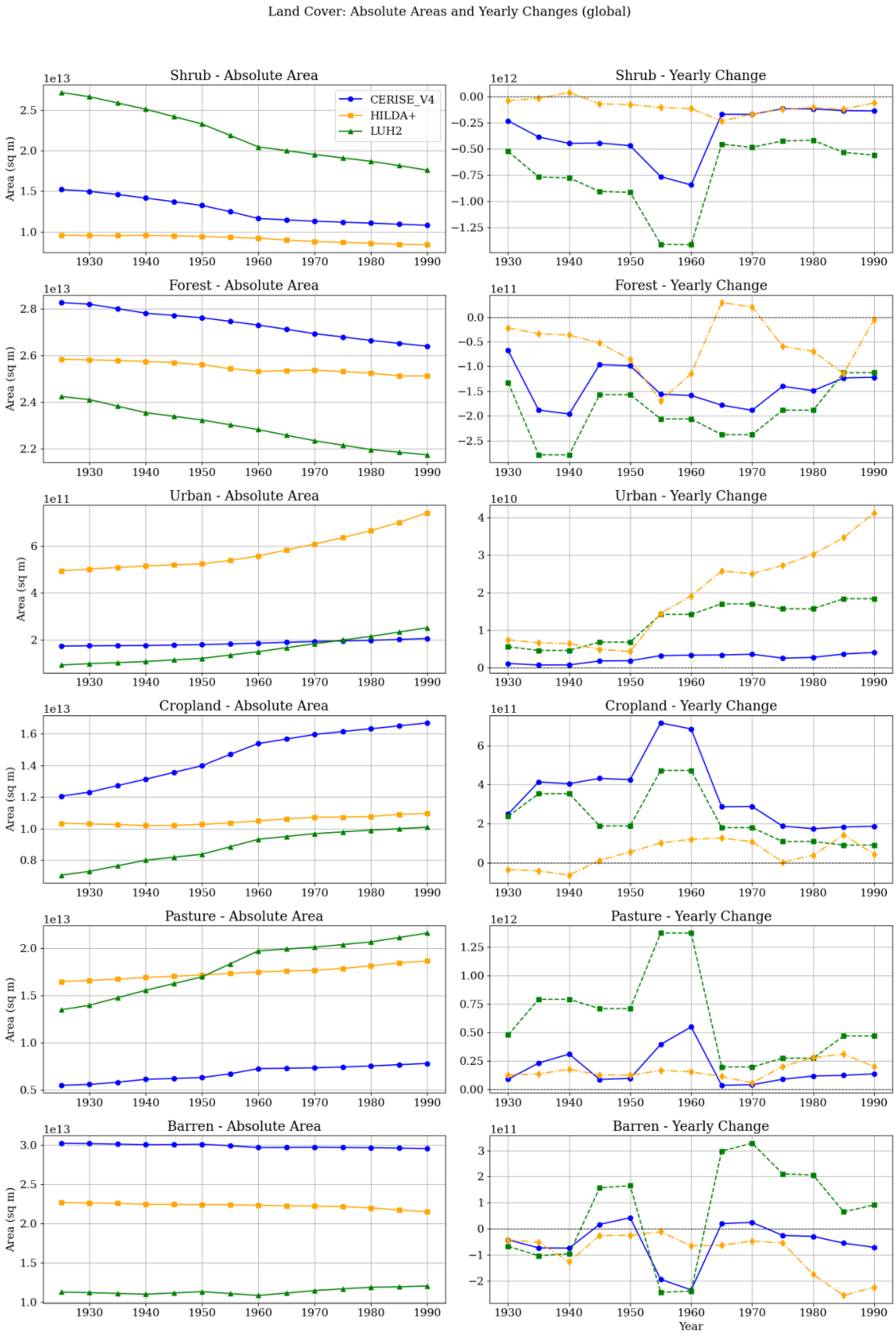

Figure 9: Comparison of basic land-use classes of HILDA+ (yellow) and LUH2 (green) along with CERISE LULC simulations (blue); absolute areas (sq. m) (1925 - 1992) and yearly differences (global scale).

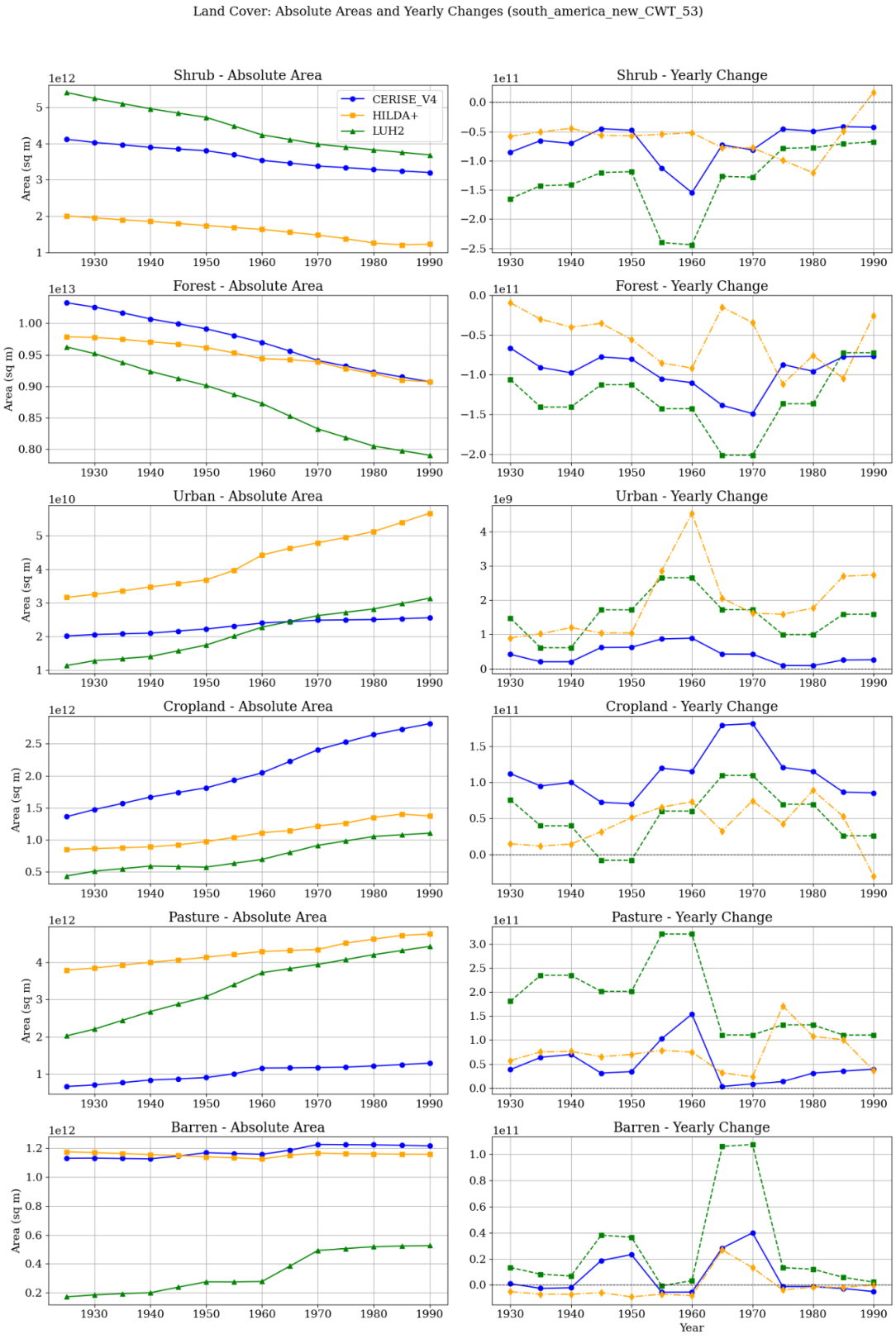

Figure 10: Absolute area and yearly difference area map over south America comparing HILDA+ (yellow) and LUH2 (green) along with CERISE LULC simulations (blue); absolute areas (sq. m) (1925 - 1992) and yearly differences (South America).

3.1.2 Spatial maps of different periods

Figures 11 and 12 present the spatial distribution of land cover types derived from the implemented classification algorithm for the South American region and the global scale, respectively. These maps highlight notable temporal changes and regional land cover dynamics between the selected timeframes.

The South American regional map (Figure 11) reveals pronounced anthropogenic impacts, with extensive deforestation patterns most evident in the Amazon Basin. These changes are especially prominent between 1925 and 1992, reflecting widespread forest clearance likely linked to agricultural expansion, logging, and infrastructure development. The spatial fragmentation and reduction in forest cover in this region underscore the intensity of human-driven land transformation processes over the study period.

In contrast, the global-scale analysis (Figure 12), one of the most prominent observations is the significant expansion and densification of shrubland across the northern African region from 1992 to 1925. Particularly in the Sahel and adjacent arid zones. This trend may be indicative of climate-driven vegetation shifts or altered land management practices. Such changes suggest either natural ecological succession or semi-arid land rehabilitation efforts in response to historic desertification. A similar trend is observed over the central United States when examining land cover transitions backward in time to 1925. Specifically, we find a notable increase in shrubland cover compared to the 1992 baseline, indicating that the LUH2 dataset attributes significant cropland-to-shrubland conversions during the early 20th century. This likely reflects land abandonment or reversion to natural vegetation prior to large-scale agricultural intensification.

Interestingly, in the southernmost regions of South America, including parts of Argentina and Chile, the land cover maps reveal a noticeable trend. Specifically, areas classified as croplands in 1992 show a notable transition to forested regions in the 1925 dataset. This suggests that reforestation or natural regeneration may have occurred in these areas post-1925, or alternatively, that agricultural land abandonment allowed for vegetative recovery over time. These findings highlight the spatial heterogeneity of land cover dynamics across different climatic and socio-political contexts within the continent.

Together, these spatial patterns emphasize the algorithm’s ability to detect both subtle and large-scale changes in vegetation cover, offering valuable insights into both climatic variability and anthropogenic land use practices across temporal and geographic scales.

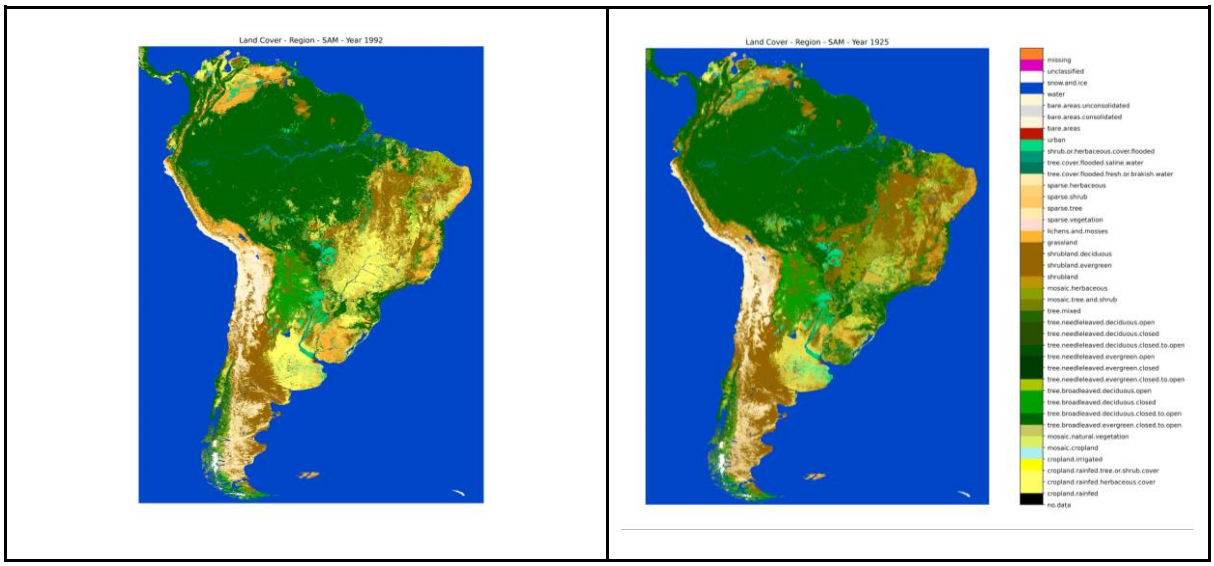

Figure 11: Historical CERISE land cover maps comparing years 1992 (left) and 1925 (right) over South America.

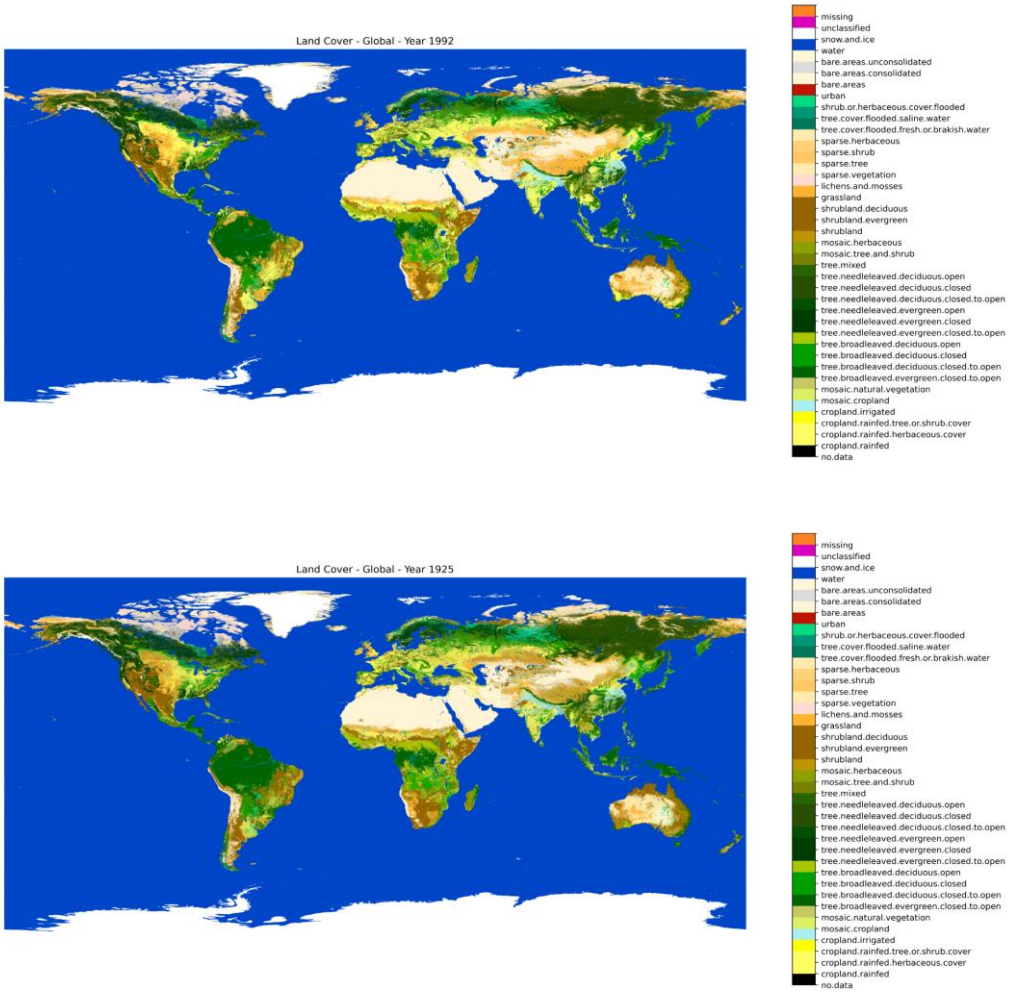

Figure 12: Global historical CERISE land cover maps comparing years 1992 (top) and 1925 (bottom).

3.2 LAI

The results of the entire pipeline are 900 files in NetCDF and Zarr covering the period from 1925 to 1999. Figure 13 illustrates a single month of the LAI ML reconstruction resultss.

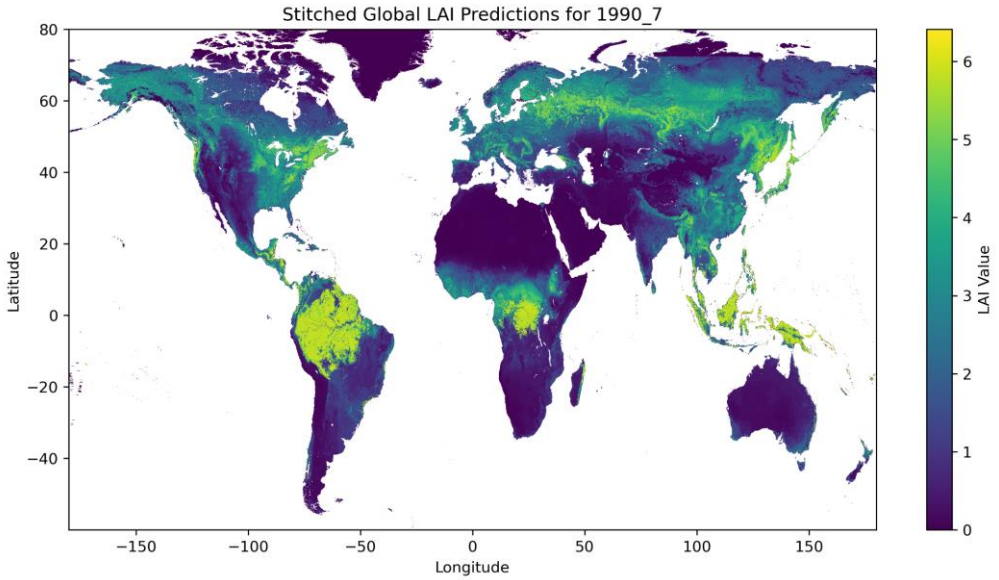

Figure 13: A global LAI reconstructed with ML for June 1990, created by stitching together regional models during post-processing.

The input data for both training and ground truth presents challenges due to its highly imbalanced nature and varying quality. The CONFESS LAI dataset is divided into two distinct periods: the years 2000-2014, which are considered reliable, and the years 1992-1999, which lack comparable quality, particularly evident in large jumps in LAI values within the Amazon region. Consequently, the earlier period (1992-1999) was excluded from training and used solely for comparison, with only the years 2000-2014 used for training and testing. The LAI values also exhibit a significant imbalance, with over 92% of global values falling within the range, potentially impacting the model's ability to capture higher values. Additionally, data quality diminishes further back in time. Both LUH2h and HILDA datasets incorporate satellite-era information from the late 20th century, but earlier years lack comprehensive global coverage. This results in the disappearance of distinctive patterns and the emergence of large patterns resembling national borders, likely due to the use of national averages in the absence of spatially accurate data. Therefore, any model trained on this dataset will inherently underperform when the less accurate early data is included. Additionally, the lack of sufficiently long, high-quality data prevents us from building a model that can accurately capture spatiotemporal relationships. As a result, we have intentionally avoided modeling the temporal aspect of the data, focusing instead on spatial relationships and relying on autoregressive prediction and the correlation between the predicament (LAI) and 15 predictors. Our approach aims to identify patterns within the spatial domain rather than establish relationships over time.

This led us to discard the years prior to 2000 for any training and testing purposes. The resulting data show continuity with the CONFESS LAI from 2000, which improves the harmonization of the CONFESS LAI for the period from 1982 to 1999 on a global scale. Figure 14 presents the global average of the monthly LAI from 1925 to 2015, for the CONFESS and

CERISE

CERISE LAI datasets in 1km resolution and AVHRR-GEO2 LAI in 4 km resolution. A time-lapse video of the reconstructed global LAI maps from 1990 to 1950 is publicly available on youtube (https://youtu.be/iZl7tIzPt-U).

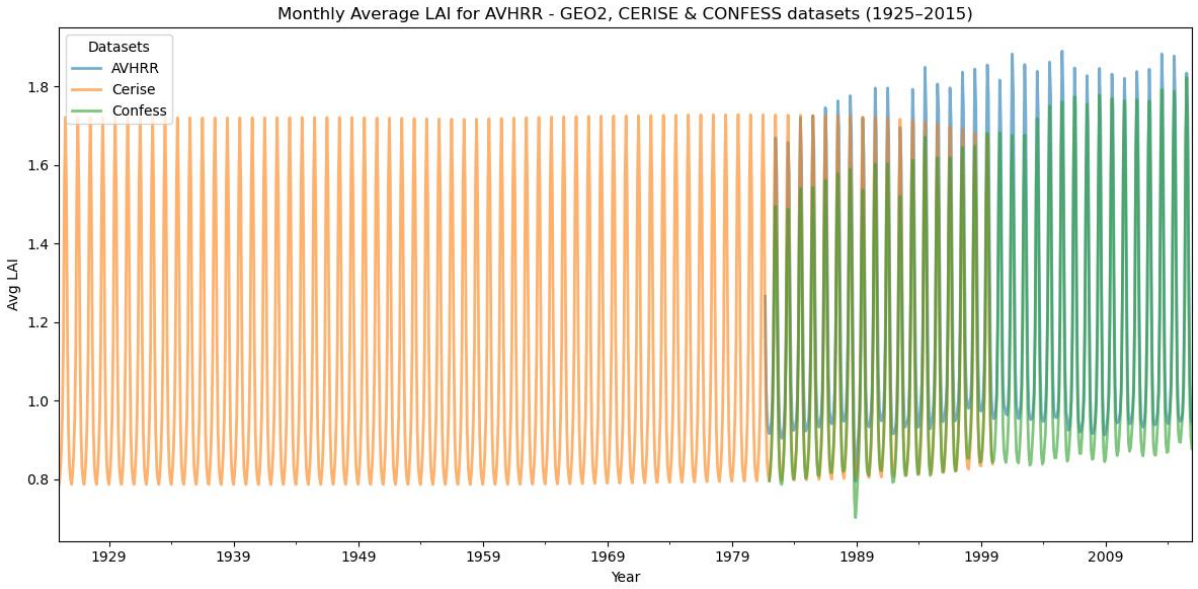

Figure 14: Time series of global average LAI values from 1925 to 2015, comparing AVHRR-GEO2 LAI (blue) in 4 km resolution, CERISE (orange) & CONFESS including the original (1993-2019) and extended CONFESS (1982-1993) (green) LAI in 1 km resolution.

As the model is trained on data from 2000 to 2014, it clearly shows a tendency to predict higher LAI values compared to the CONFESS LAI. This tendency persists further into the past as the inference continues.

One known issue with CONFESS LAI was that in areas with high LAI, especially in the Amazon forest, there are strong discontinuities in LAI values and trends in the years prior to 2000 due to the inherent differences in the datasets that constitute the overall dataset. This issue is visible in Figure 15, bottom panel, examples 3 and 6. Despite the improvement in achieving a smoother transition to earlier years at the global scale, the performance varies significantly depending on the study region. We compared the changes across several areas with high LAI values and observed differing behaviors in smaller regions. Figure 15 illustrates how LAI changes across four study areas, marked on the accompanying maps. The CONFESS LAI is shown in green, while the CERISE LAI is displayed in orange for the overlapping period and in blue for the remaining years. In regions with relatively low LAI, like most of the Northern Hemisphere, we see an increase in LAI values as backward temporal reconstruction progresses from 1990 towards 1925. In contrast, in areas with high LAI values (such as the Amazon, central Africa, and East Asia), we see that LAI values are decreasing as the reconstruction progresses towards the year 1925.

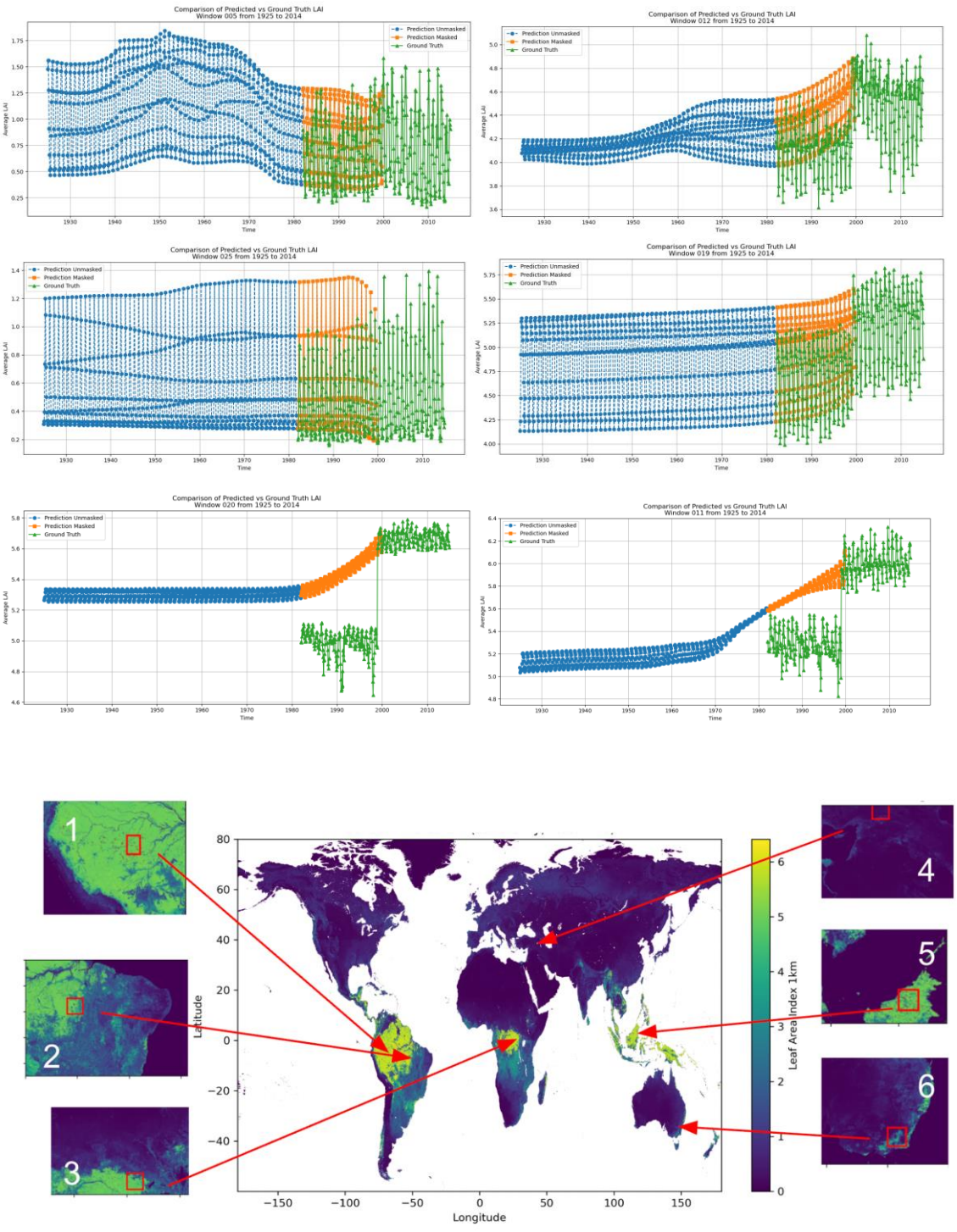

Figure 15: Time series of regional average LAI values from 1925 to 2000, comparing CONFESS LAI (ground truth in green) and CERISE LAI (blue and orange).

We believe this results from the model's training objective to maximize the fit to the average LAI. This objective tends to incentivize the model to reduce extreme high and low predictions, pulling estimates closer to the average values within the training regions. Unfortunately, this leads the model to either significantly underestimate or overestimate LAI values in certain regions, even while achieving a good fit to the overall average LAI. Figure 16 below shows the percentage difference in the time-varying CERISE high LAI relative to the IFS climatological high LAI (climate V21). The top left plot shows that in January 1940, there was about a 40%

reduction in CERISE high LAI over much of the southern hemisphere compared with climate V21. In contrast, the northern hemisphere shows a 40% increase in high LAI widely over North America and Europe. These regional differences appear to be overestimated, even if the global average differences are relatively small. Furthermore, the sharp reduction in high LAI over South America seems counter-intuitive, given that high vegetation cover was much larger in the CERISE dataset for 1940 than in climate V21. The equivalent plot for July 1940 is shown in the top right and shows different patterns in the northern hemisphere compared to January 1940, with a 40% reduction in the CERISE high LAI over much of North America and Europe. A reduced signal is present over the southern hemisphere relative to January 1940. The equivalent plots for January and July 1960 are shown below the 1940 plots. The plots for 1940 and 1960 are almost identical, demonstrating very limited inter-annual variability in the CERISE LAI datasets between 1940 and 1960. This is expected given that the training used to generate the CERISE LAI datasets does not take into account the seasonal weather anomalies.

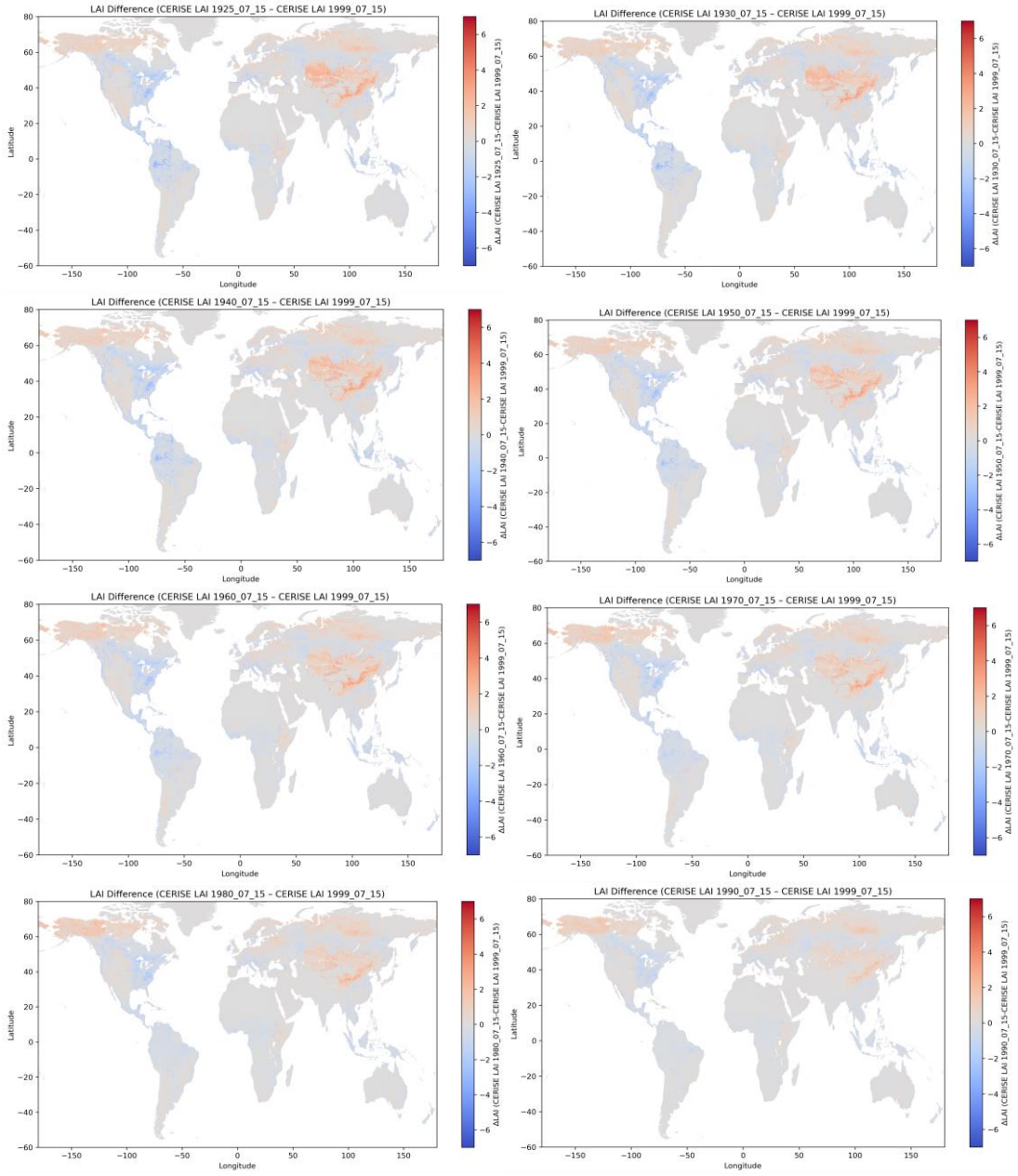

Figure 16: Figures illustrate the Leaf Area Index (LAI) difference between several past years (1925, 1930, 1940, 1950, 1960, 1970, 1980, 1990) and 1999. Red areas indicate higher LAI values in the earlier year, while blue areas indicate higher LAI in 1999.

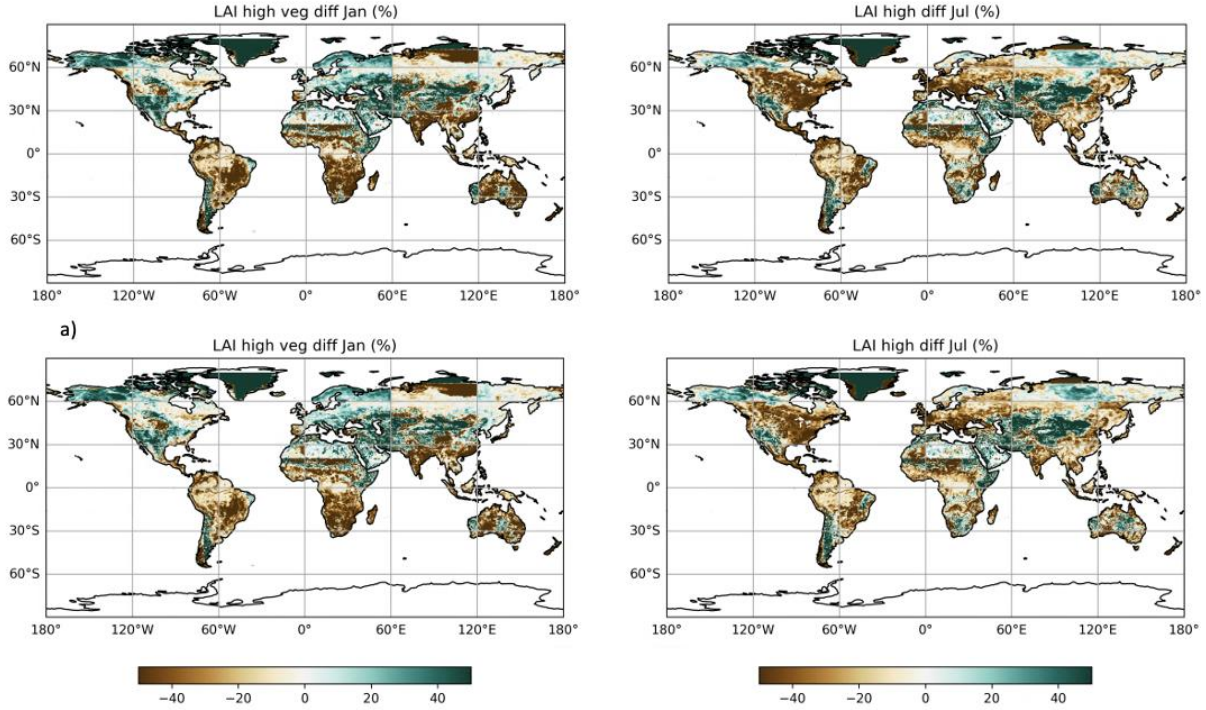

Figure 17: Percentage difference for CERISE high LAI vs climate V21 high LAI (positive means more LAI in CERISE). The top left (right) plot shows the % difference for January (July) 1940 and the bottom plots are the equivalent for 1960.

In addition to the general tendency of the model to shift LAI predictions toward regional averages, as discussed above, another contributing issue we have observed is the model's sensitivity to significant changes in the input data. These shifts, which often occur on a decadal scale, are reflected in the model's predictions.

3.3 Lake Cover

Global seasonally varying water distribution maps were generated based on high horizontal (30 meters) and temporal (month) resolution satellite data for the past 50 years and some high-fidelity auxiliary data (e.g. coastline shapefiles, elevation datasets). The main data input is from the Joint Research Centre (JRC) Global Surface Water Explorer (GSWE) dataset (Pekel et al., 2016) at 30 meter resolution (grid EPSG:4326) covering the period 1984-2021: 'monthlyHistory' – monthly water distribution with water/ notWater/ noData data, to further customise water classes based on a relevant period (e.g. 2012-2021). In addition, are used Copernicus GLO30 dataset at 30 meter resolution (grid EPSG:4326) 'waterBodyMask' to separate the ocean from inland water; and numerous regional glacier datasets at 15-100 m resolution (i.e. British Antarctic Survey, QUANTARCTICA, GIMP project, QGREENLAND, Norwegian Institute, Icelandic Met service) to improve water distribution over relevant regions.

Global seasonally varying water distribution maps generated for 1992-2021 are fully independent and are purely based on satellite data. Earlier (1962-1991) maps have in general the 1992-2001 period as a baseline and are updated only regionally - based on available

reliable satellite information or historic records (i.e. maps, verbal description) with supplementary elevation data criteria.

Generated maps are grouped per 10-year periods (all available periods: 1962-1971, 1972-1981, 1982-1991, 1992-2001, 2002-2011, 2012-2021), each decade has one permanent water map and twelve monthly maps (i.e. permanent water + monthly delta):

- 1992-2021 (i.e. 1992-2001, 2002-2011, 2012-2021) maps are fully independent and are purely based on satellite data;
- 1962-1991 (i.e. 1962-1971, 1972-1981, 1982-1991) maps have in general the 1992-2001 period as a baseline and are updated only regionally - based on available reliable satellite information or historic records (i.e. maps, verbal description), and elevation data criteria;
- 1925-1961 (i.e. 1925-1931, 1932-1941, 1942-1951, 1952-1961) maps use the maps from the 1962-1971 decade, because (i) it had booming water-related anthropogenic activities, i.e. building of large reservoirs and irrigation channels, re-routing rivers, etc.; (ii) it was the last decade with the acceptable amount and quality of in situ data to make any assumptions/ calculations or verification.

The regional map corrections and updates were implemented for regions with:

- frozen 2012-2021 distribution – north of 78°N, Sao Tome island and south-west Africa, Antarctica and South Georgia and the South Sandwich Islands;
- 1982-1991 baseline distribution – large regions surrounding Great Salt Lake (USA), Mississippi River (USA), Brazil, Argentina, Australia;
- altered baseline distribution to match reality – Toshka lakes (Egypt, formed in 1998), Dubai islands (UAE, built in 2001-2003), Great Salt Lake (USA, 1962-1971 water distribution mimics 2012-2021 period);
- updated baseline distribution to match historical information – Great Salt Lake (USA), Lake Mead (USA), Lake Poopo (Bolivia), Dead Sea (Israel/Jordan), Lake Habbaniyah (Iraq), Lake Milh (Iraq), Lake Urmia (Iran), Aral Sea (Kazakhstan/Uzbekistan), Menindee Lakes (Australia), Lake Chad (Chad/Cameroon/Nigeria/Niger).

Different examples for Aral Sea (Kazakhstan/Uzbekistan), Toshka Lakes (Egypt), and Tonle Sap Lake (Cambodia) regions are shown in Figures 18, 19, and 20 respectively. Each example has permanent water map changes throughout six decades, inland water cover currently used operationally (climate.v021) for IFS model at ECMWF, recent satellite image for reference, and a histogram of total water area over region in sq.km per each month.

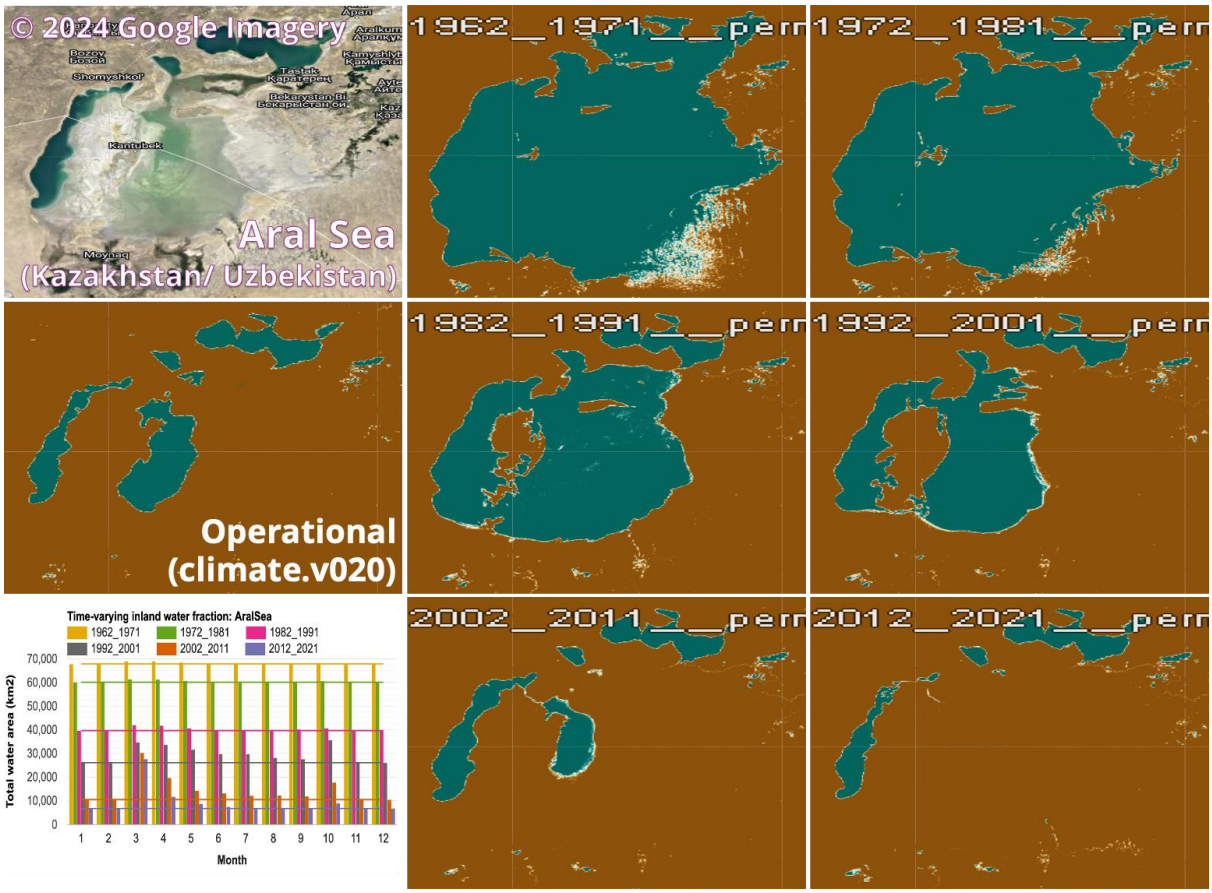

Figure 18 - Permanent water distribution for operational static and time-varying decadal maps over the Aral Sea region in Kazakhstan/Uzbekistan, in the lower left corner the histogram shows the total water area over the region in sq.km per month.

The Aral Sea (see Figure 18) had a massive diversion of water for cotton and rice cultivation in the 1960's after which rapid shrinking of it began. After collaborative restoration works, the surface area of the Northern part became stable and for the Western and Eastern parts shrinking slowed significantly (middle and right column plots of Figure 18 correlate with these events). The current static operational water distribution map (see left column middle row of Figure 18) represents well the average of monthly maps for the 2012-2021 period (see left column bottom row of Figure 18), but has strong underestimation compared to water distribution from previous decades.

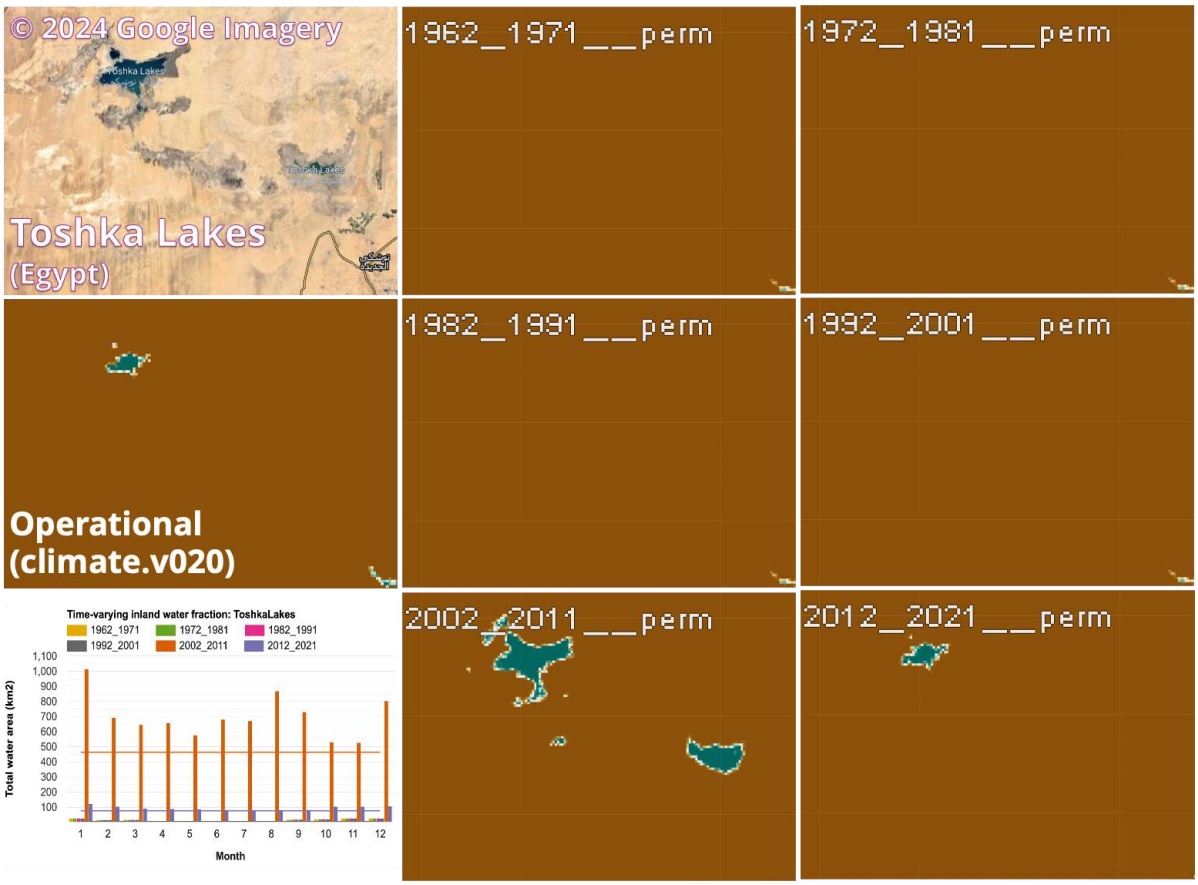

Figure 19 - Permanent water distribution for operational static and time-varying decadal maps over the Toshka Lakes region in Egypt, in the lower left corner histogram shows the total water area over the region in sq.km per month.

The Toshka Lakes (see Figure 19) only formed in 1998 due to massive flash floods and river floods in Ethiopia, which caused floodwaters to flow down the Nile River. Formed lakes first boomed the agriculture in the region, but soon started drying, and became almost empty by 2018 (middle and right column plots of Figure 19 correlate with these events). The current static operational water distribution map (see left column middle row of Figure 19)represents well permanent water for the 2012-2021 period (see right column bottom row of Figure 19), but has strong underestimation comparing to water distribution from the previous decade 2002-2011, and overestimation comparing to even earlier decades (region was a desert). Due to annual heavy rainfall and major flooding events in Sudan and South Sudan, since 2018 the Toshka Lakes started refilling and even formed one new lake - this regrowth was not captured by the data as maps represent the best suited monthly distribution over 10-year period (see left column bottom row of Figure 19).

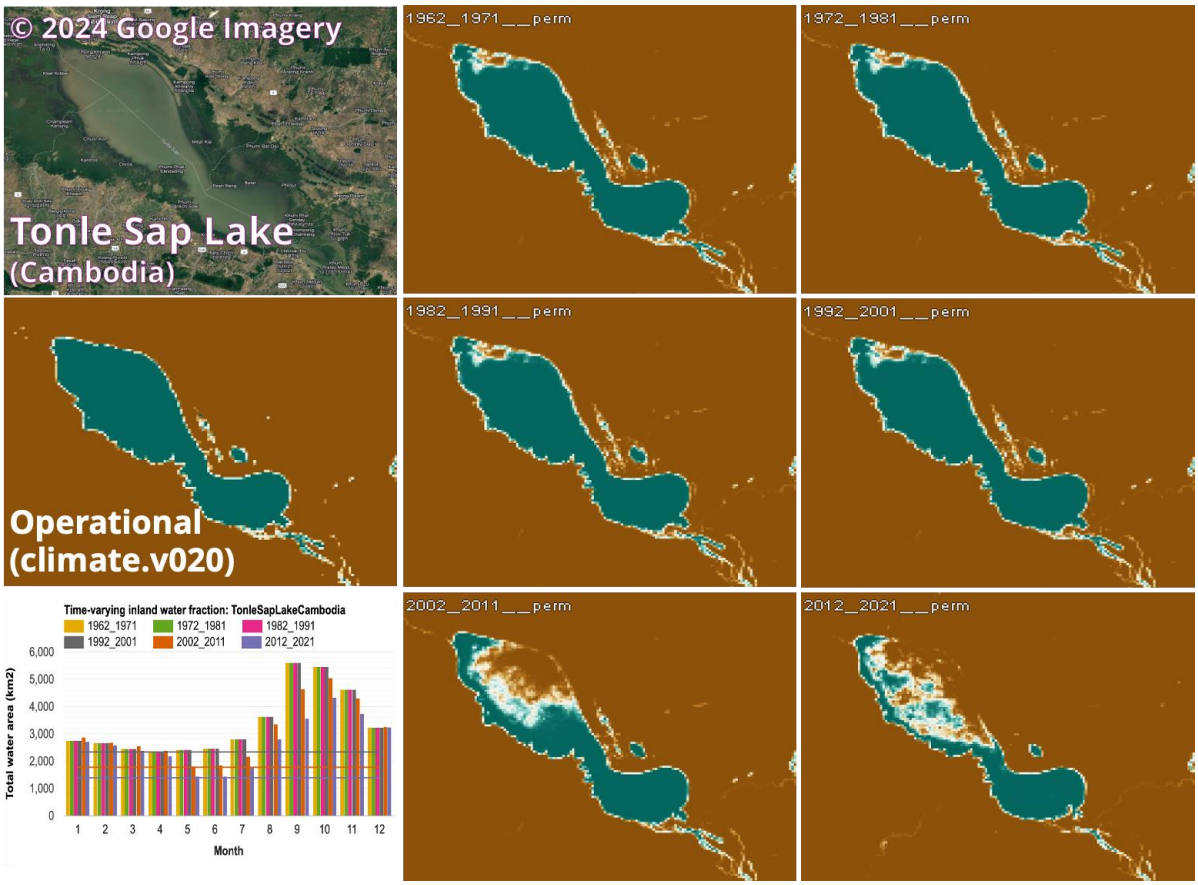

Figure 20 - Permanent water distribution for operational static and time-varying decadal maps over the Tonle Sap Lake region in Cambodia, in the lower left corner the histogram shows the total water area over the region in sq.km per month.

The Tonle Sap Lake's (see Figure 20) water level started lowering down in the last few decades due to dam construction in the tributaries of the Mekong river, and lowering was enhanced in the last years (especially in 2020) by El Nino (middle and right column plots of Figure 20 correlate with these events). The current static operational water distribution map (see left column middle row of Figure 20) represents well the average of monthly maps for 2012-2021 period (see left column bottom row of Figure 20), but shows some underestimation compared to yearly average water distribution from previous decades. The yearly cycle of the lake's water cover is well captured by the data and follows rainy (from May to October) and dry (from November to March) seasons (see left column bottom row of Figure 20).

3.3.1 Direct evaluation

Generated maps were aggregated in space and time to understand if patterns of inter-annual and seasonal variability globally and regionally follow known evolution. While some first examples were discussed in the section above, in this section we will see direct regional comparison in more detail.

Over the Northern hemisphere (90°N - 20°N, see top row of Figure 21) the total water area curve (see right column top row of Figure 21, yellow line) shows slight decreasing trend over decades, which could be explained (especially in recent years, 2012-2021) by the decrease of small water bodies (see middle column top row of Figure 21, fractions from 0.0 to 0.5). Total water area has a clear yearly cycle (see right column top row of Figure 21), which can be explained in general for some regions by high-pressure systems and cold air masses in the

winter leading to drier conditions and reduced surface water distribution, and by monsoons in the summer leading to intense rainfall and increased surface water distribution.

Over the tropics (20°N - 20°S, see middle row of Figure 21), there is a total water area decrease from 1962-1971, which can be explained by Chad Lake shrinking (currently by ~90%) due to variations in climate and water withdrawals for irrigation. Total water area has no pronounced yearly cycle (see right column middle row of Figure 21).

Over the Southern hemisphere (20°S - 90°S, see bottom row of Figure 21), the total water area curve (see right column bottom row of Figure 21, yellow line) shows a decreasing trend. For recent years (2012-2021), this could be explained by the fact that several big lakes and reservoirs shrunk or became seasonal (i.e. non-permanent) due to water use for irrigation and lack of rainfall, or due to glacier melt, which reduces water inflow into glacier-fed water bodies. Total water area has a slightly more pronounced yearly cycle than in the tropics (see right column bottom row of Figure 21). Since some parts of the Southern hemisphere have two rainy seasons while other parts have only one, the cumulative signal could be mixed. The current plot depicts e.g. the rainy season in Southern Australia.

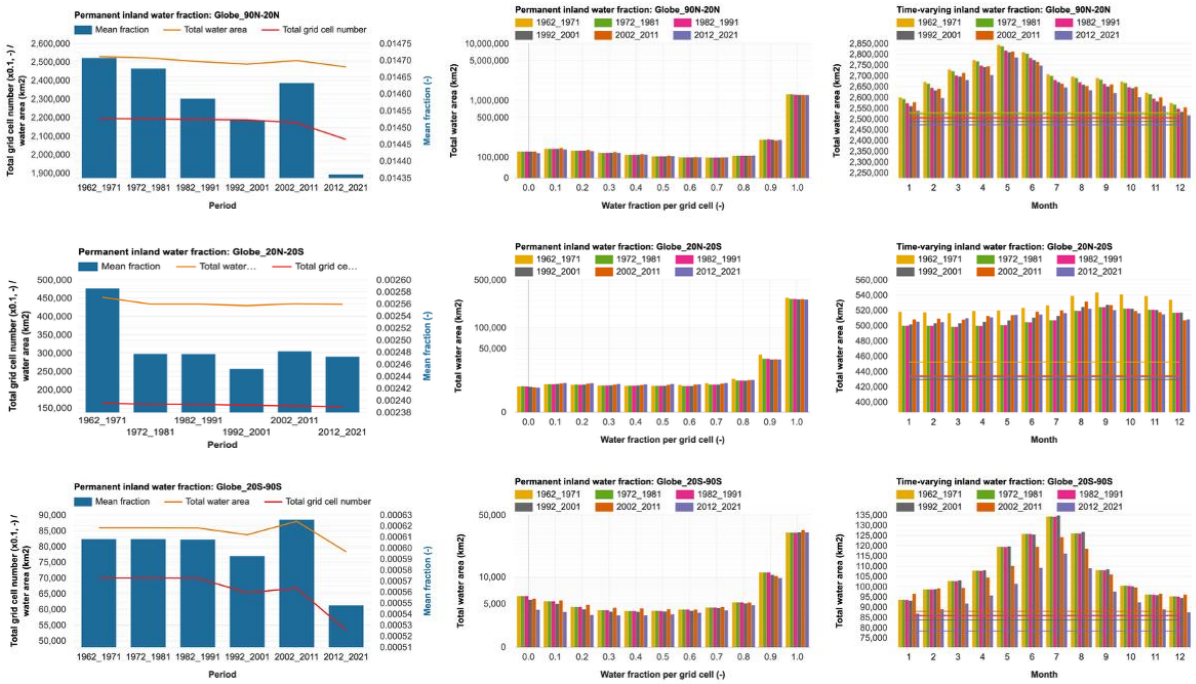

Figure 21: Permanent inland water mean fraction, total grid cell number per decade and total water area in sq.km shown: per decade only (left), per grid cell water fraction and decade for permanent water (middle) and per month and decade for time-varying water (right). Statistics are produced from 1km resolution data for regions 90°N - 20°N (top row), 20°N - 20°S (middle row), 20°S - 90°S (bottom row).

Over Europe (see Figure 22), the total permanent water cover is rather stable from decade to decade, as most lakes are permanent and situated in the Boreal climate zone, with a seasonal cycle that follows the period of increased precipitation.

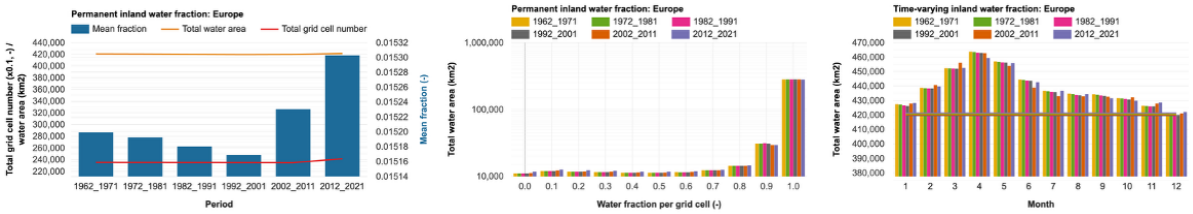

Figure 22 - Permanent water mean fraction and total grid cell number (per decade), total water area in sq.km (per decade only, and per grid cell water fraction and decade for permanent water; per month and decade for time-varying water) at 1km resolution for the European region.

Examples of regional map evaluation:

- Over the Aral Sea (see Figure 23), the total permanent water cover is decreasing and monthly water cover seasonality is increasing (i.e. difference between min and max of 12 months), which is supported by the historical drying and shallowing trend of the water body;
- Over the Poopo Lake (see Figure 24), the total permanent water cover changes from decade to decade following major drought events, and recent decades decrease is associated with climate change (i.e. melt of Andes glaciers and increased draughts) and continuous water diversion for mining and agriculture;
- Over the Lake Urmia (see Figure 25), the total permanent water cover is decreasing from decade to decade and the decrease is more rapid in recent years. This is supported by Iran's decision to grow irrigation-intensive crops since 1979. Monthly water cover seasonality is increasing (i.e. difference between min and max of 12 months) as the lake becomes shallower, and water monthly cover follows rainy seasons occuring from January till March and from October till December.
- Over the Toshka Lakes (see Figure 26), the total permanent water cover follows historical formation and captures seasonality well. Maps do not capture recent re-filling of the lakes as the used dataset lags 3 years (currently available till end of December 2021).

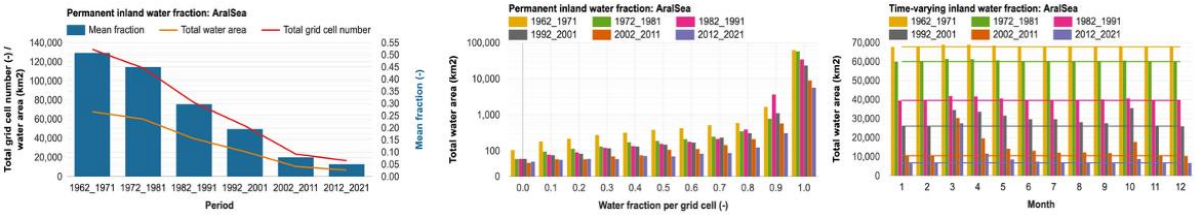

Figure 23 - Permanent water mean fraction and total grid cell number (per decade), total water area in sq.km (per decade only, and per grid cell water fraction and decade for permanent water; per month and decade for time-varying water) at 1km resolution for the Aral Sea region.

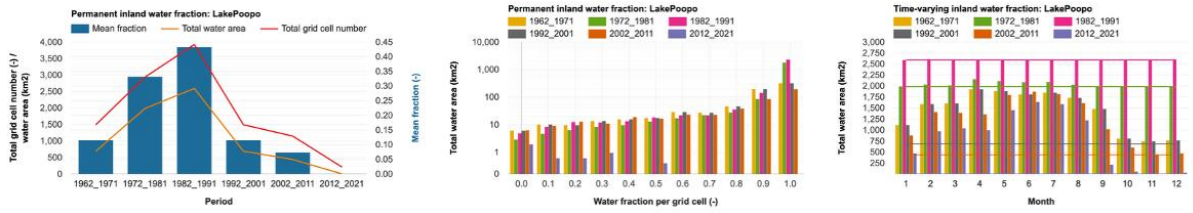

Figure 24 - Permanent water mean fraction and total grid cell number (per decade), total water area in sq.km (per decade only, and per grid cell water fraction and decade for permanent water; per month and decade for time-varying water) at 1km resolution for the Poopo Lake region.

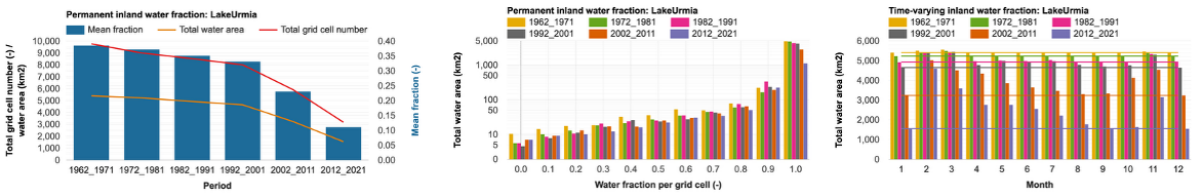

Figure 25 - Permanent water mean fraction and total grid cell number (per decade), total water area in sq.km (per decade only, and per grid cell water fraction and decade for permanent water; per month and decade for time-varying water) at 1km resolution for the Urmia Lake region.

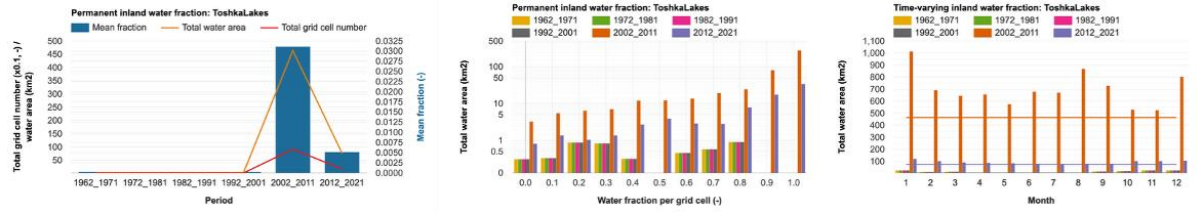

Figure 26 - Permanent water mean fraction and total grid cell number (per decade), total water area in sq.km (per decade only, and per grid cell water fraction and decade for permanent water; per month and decade for time-varying water) at 1 km resolution for the Toshka Lakes region.

Generated maps were also compared with available reliable global datasets according to the total water area in sq.km at 1 km resolution (i.e. fraction at 1 km resolution multiplied by area of a 1 km resolution grid cell, and with all values over the region of interest are summed up):

- ECMWF_perm|seasMean|seasMax (current) - generated maps were aggregated globally and regionally to compare min (i.e. only permanent water, always present), mean (i.e. averaged over 12 months), and max (i.e. maximum over 12 months) water extents; fractional information at 891 m near Equator resolution is derived from 30-meter discrete information;
- ESACCI_water - ESA CCI yearly maps represent maximum yearly water extent; discrete information at 300 m near Equator (i.e. 10 arc sec);
- CGLS_perm|seas - CGLS yearly maps represent min (i.e. only permanent water, always present) and max (i.e. maximum of the year) water extents; fractional information at 100 m near Equator resolution;
- WorldCover_water - ESA WorldCover yearly maps represent (on average) maximum yearly water extent; discrete information at 10 m near Equator.

Over the Northern hemisphere (90°N - 20°N), the generated maps show consistently less water than in ESA CCI (see top row of Figure 27). The most probable reasons for that are: (i) difference in nominal data resolution (i.e. 30 m vs 300 m for ESA CCI); (ii) the difference in water type, where the generated maps capture monthly variations with the best estimate over

the decade being selected, while ESA CCI captures the maximum water extent of the specific year (ocean filtered with specially generated constant in time mask for 2015).

Over the Tropics (20°N - 20°S; see middle row of Figure 27), the generated maps show a close match between ECMWF_seasMean and ECMWF_seasMax, which means that water is present there most of the year yet not permanently. Comparison with ESA CCI shows exactly the same as for Northern hemisphere and for the same reasons. We have also compared generated maps with CGLS and WorldCover and all of them compare very well (i.e. total amount of water is preserved which is important).

Over the Southern hemisphere (20°S - 90°S), the generated maps show good correlation of ECMWF_seasMax and ESA CCI (see bottom row of Figure 27), with slight overestimation (i.e. less than 5’000 sq.km) in early 1990’s and underestimation (i.e. less than 12’000 sq.km) after 2002. We have also compared generated maps with CGLS and WorldCover. CGLS has a strong overestimation of water extent into ocean, both for permanent and seasonal water over the southern part of Chile (area of and around Parque Nacional Bernardo O'Higgins). Most probably due to this overestimation, the total area of permanent water in CGLS is consistent with the maximum water extent from WorldCover, ESA CCI and ECMWF_seasMax and CGLS_seas is almost double comparing to other datasets. The WorldCover data is almost identical to ECMWF_seasMax, and the total water area is slightly lower than ESA CCI, which can be explained by the difference in their native resolutions (i.e. 10 m WorldCover vs 300 m ESA CCI).

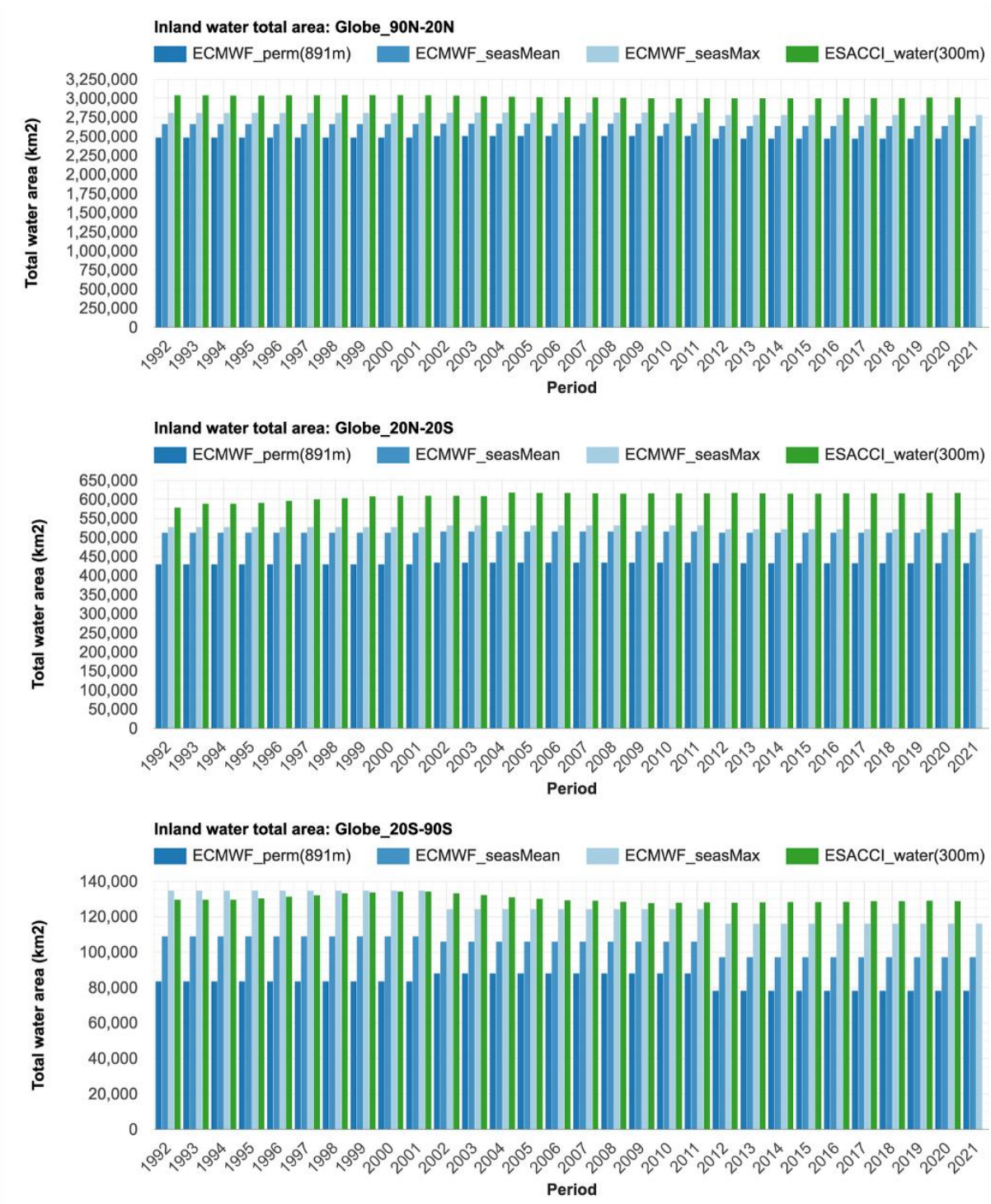

Figure 27 - Dataset's total water area in sq.km comparison at 1 km resolution for 90°N-20°N (top row), 20°N-20°S (middle row), and 20°S-90°S (bottom row); ECMWF_perm refers to permanent (i.e. always present) water aggregated from generated maps, ECMWF_seasMean - averaged over 12 months water, ECMWF_seasMax - maximum over 12 months, ESACCI_water - yearly maximum water extent from ESA CCI data.

Over Europe (see top row of Figure 28), the comparison of the generated ECMWF_seasMax with ESA CCI, CGLS_seas, and WorldCover shows consistently slightly less water in ECMWF_seasMax, which can be also explained by the difference in the initial datasets'

resolution and water type. All three datasets have a perfect match within them. ECMWF_perm is almost identical to CGLS_perm. This confirms that generated maps conserve the total water budget over Europe.

Over the Aral Sea in Kazakhstan/ Uzbekistan (see middle row of Figure 28), the generated maps have a very good correlation with other datasets (e.g. 1992-2001 ECMWF_perm value is the same as ESA CCI data for 2001 and ECMWF_seasMax is almost the same as ESA CCI data for 1992). The main point to remember is that ECMWF data represents the best fit over the whole 10-year period, while other datasets are yearly.

Over the Toshka Lakes in Egypt (see bottom row of Figure 28), the generated maps and ESA CCI have good agreement on capturing the creation of the lake, and together with CGLS all three datasets perfectly follow the lake's drying period. ESA CCI, CGLS, and WorldCover depict well the refilling of the lake, while the generated maps miss it completely as they used data only up to 2021 where the last years of re-filling were considered as outliers for the 10-year period.

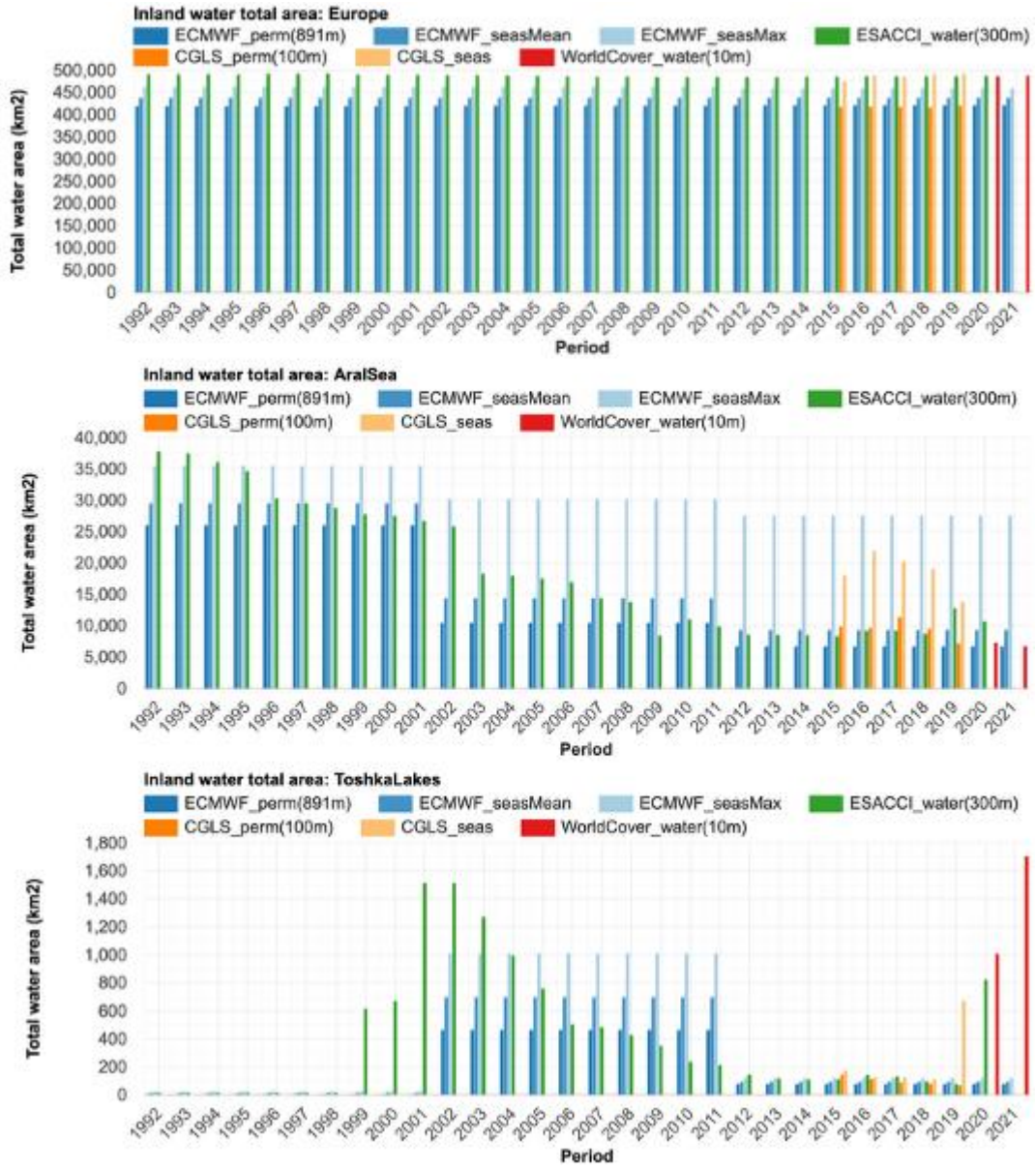

Figure 28 - Dataset's total water area (in sq.km) comparison at 1 km resolution for Europe (top row), Aral Sea (middle row), and Toshka lakes (bottom row); ECMWF_perm refers to permanent (i.e. always present) water aggregated from generated maps, ECMWF_seasMean - averaged over 12 months water, ECMWF_seasMax - maximum over 12 months, ESACCI_water - yearly maximum water extent from ESA CCI data.

In general, the generated maps have a good coherence for permanent and maximum of the year water distribution. Comparison with ESA CCI generally indicates that generated maps underestimate water cover due to the initial resolution and water type. Nevertheless, sometimes coarser resolution maps with discrete information might underestimate water bodies extent with irregular coasts, e.g. over Lake Chad in Chad/Cameroon/Nigeria/Niger (see top row of Figure 29). Comparison with ESA CCI constantly shows that even ECMWF_seasMax is underestimating water cover, yet comparison with much higher resolution WorldCover data shows good coherence. Sore areas are very seasonal and weather dependent, e.g. the Poopo Lake in Bolivia (see bottom row of Figure 29), where all datasets show rather different results, nevertheless ECMWF_perm coincides with recorded transition of the lake into a salt pan due to no more water coming from the melted glaciers in Andes and continuous water diversion for mining and agriculture.

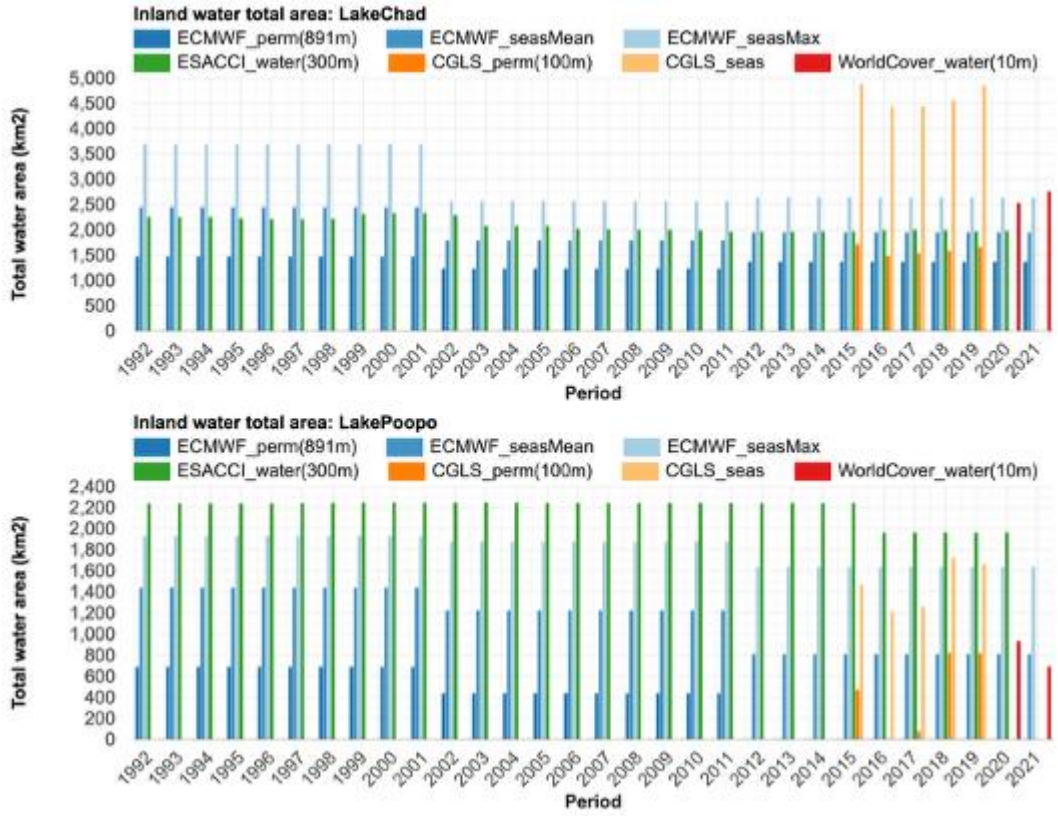

Figure 29 - Dataset's total water area (in sq.km) comparison at 1 km resolution for Lake Chad (top row), and Lake Poopo (bottom row); ECMWF_perm refers to permanent (i.e. always present) water aggregated from generated maps, ECMWF_seasMean - averaged over 12 months water, ECMWF_seasMax - maximum over 12 months, ESACCI_water - yearly maximum water extent from ESA CCI data.

3.3.2 Indirect evaluation

Generated maps were also evaluated indirectly by running a numerical offline (i.e. no feedback to the atmosphere) open-loop (i.e. no data assimilation) experiment with the IFS model (CY49R1) at ~25 km resolution (Tco399) over 1995-2019, and comparing results (i.e. skin temperature) with the high fidelity satellite composite product of skin temperatures CCI LAKES (nominal resolution 1 km, represent ~10.30 am local time; Carrea et al., 2024). Figure 30 shows a recent satellite image and the Aral Sea representation in the current operational version of IFS at ~25 km resolution, as well as January, April, July and October months from time-varying lake covers over three recent decades (i.e. 1992-2001, 2002-2011, and 2012-2021).

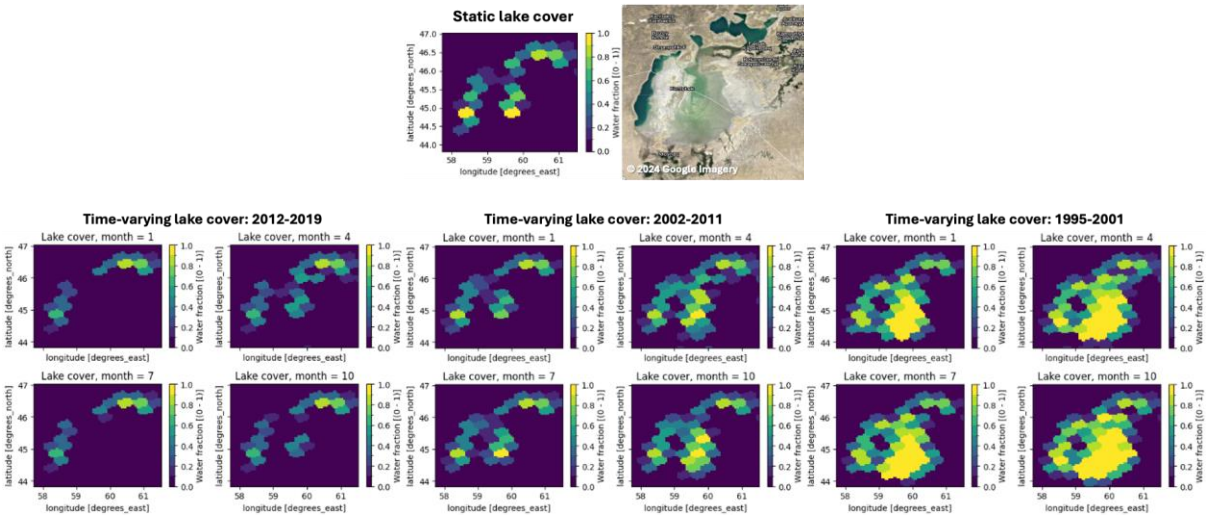

Figure 30 - Static (current operational IFS model's climate.v021, top row) and time-varying (for 1995-2001, 2002-2011, 2012-2021 decades, for January, April, July and October months, two bottom rows) lake covers at ~25 km resolution (Tco399) for the Aral Sea region.

The IFS model's offline version was adapted to use monthly lake and land sea covers and to update soil moisture and soil temperature based on lake filling or drying.

Comparisons with skin temperature observations revealed, e.g. over the Aral Sea area an overestimation of water cover for March in 2002-2011 and 2012-2019. This is due to the limited number of valid satellite observations available to generate maps for March (e.g. datasets used produced by JRC use satellite image's visual diapason to identify water surface, which is challenging during dark winter months or cloudy weather). It was revealed that due to a missing lake salinity parameter in the model - which results in earlier ice-on and later ice-off dates, and due to underestimation of the mixed-layer depth by the model - it results in a mixed-layer temperature overestimation in summer (see left and middle columns top row of Figure 31).

In 1995-2001 the Aral Sea water cover is much wider and match that time cover much better than the static (i.e. represents ~2018 year) lake cover. Model results still have overestimation of the duration of the ice-on period due to missing lake salinity parameter, but mixed-layer temperature overestimation in summer becomes negligible (see right column top row of Figure 31).

Even though single current static lake cover represents quite well lake cover over 2012-2019, use of monthly maps gives on average 1.0 K reduction in yearly bias (see left column bottom row of Figure 31) and 0.2 K reduction in yearly RMSE. Historically the Aral Sea had bigger

water area in 2002-2011 than the current period. More realistic water representation for that period gives on average 2.5 K reduction in yearly bias (middle column bottom row of Figure 31) and 3.0 K reduction in yearly RMSE. In 1995-2001 the Aral Sea area was much bigger than in present time and much more stable during the year (see right column of Figure 31) as the water body on average was deeper. Realistic water representation for that period gives on average 4.5 K (up to 6.0 K in 1995) reduction in yearly bias (see middle column bottom row of Figure 31) and 4.5 K reduction in yearly RMSE.

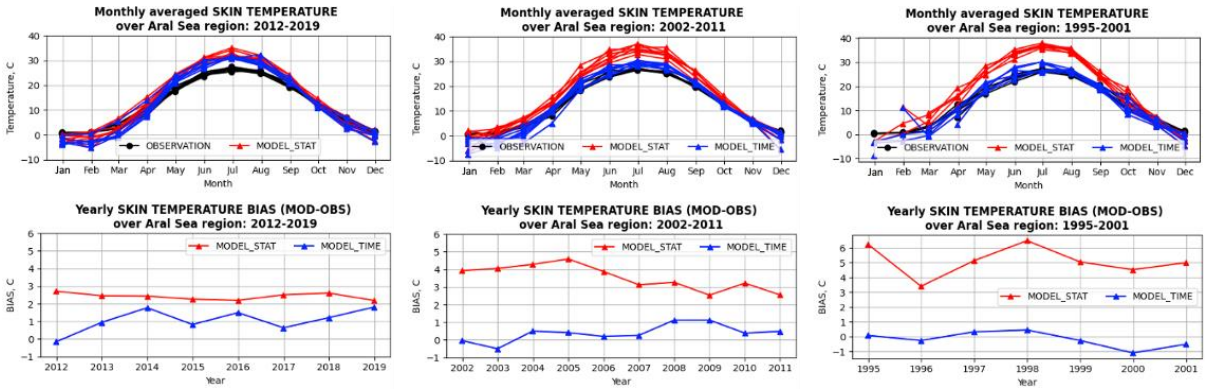

Figure 31 - Top row: Monthly and regionally averaged skin temperature yearly cycles (10 years), for observations from CCI LAKES (OBSERVATION, black color), operational IFS model that uses single static lake cover (MODEL_STAT, red color), and adapted IFS model that uses monthly lake covers (MODEL_TIME, blue color); bottom row: yearly and regionally averaged skin temperature BIAS (MODEL - OBS) over the Aral Sea region for the 1995-2001 (right column), 2002-2011 (middle column), and 2012-2019 (left column) periods.

4 Conclusion

4.1 Land cover

- **What has been achieved?**

In this study, we present a novel approach for generating yearly historical land cover maps spanning the period 1925 to 1992, leveraging an efficient, parallelized algorithm that integrates LUH2 transition data with the 1992 ESACCI land cover baseline. By systematically applying the transition information in reverse, the algorithm successfully reconstructs consistent global land cover maps that align with the ESACCI classification scheme, effectively extending its spatial-temporal coverage back in time.

The algorithm demonstrates strong potential in harmonizing heterogeneous datasets particularly reconciling the LUH2 land-use transitions with the high-resolution satellite-derived ESACCI maps. This harmonization ensures that the historical reconstructions remain consistent with modern observations, while still reflecting the temporal evolution of land cover as captured by the LUH2 framework.

- **What are the shortcomings ?**

Despite its effectiveness, the algorithm faces few limitations:

- Dependence on LUH2 Transition Data: The accuracy and fidelity of historical reconstructions are inherently limited by the quality, resolution, and completeness of the LUH2 transition data. LUH2 does not capture all nuanced local-scale land cover dynamics.

- Validation Challenges: Long-term, high-resolution historical land cover datasets are scarce, making quantitative validation difficult. Although comparisons were made with existing datasets such as LUH2 and HILDA+, notable inconsistencies emerged, partly due to methodological and definitional differences.

- Spatial Artifacts: The algorithm operates within fixed LUH2 subgrid boundaries, resulting in spatial artifacts such as blockiness in the output maps. These artifacts reduce the realism and continuity of simulated land cover.

- Unmapped Transitions: In cases where LUH2 specifies transitions not represented in the ESACCI baseline, the algorithm resorts to default fallback mechanisms. While necessary, these heuristics introduce additional uncertainty into the historical reconstructions.

- **Future work**

To address the current limitations and improve the robustness and realism of historical land cover simulations, future work should consider the following enhancements:

- Smoother Spatial Transitions: Incorporating spatial interpolation or machine learning-based spatial regularization techniques could reduce subgrid-level artifacts and produce more spatially continuous land cover transitions.

- Enhanced Data Integration: Future versions of the algorithm could integrate alternative satellite-derived datasets, historical aerial imagery, or ancillary sources (e.g., historical

maps, land surveys) to better inform land cover assignments during ambiguous transitions.

- Probabilistic or Rule-based Assignment: Replace fallback heuristics with probabilistic modeling or rule-based systems informed by ecological plausibility and regional land use practices to better resolve cases where LUH2 transitions do not align with ESACCI categories.

- Cross-Dataset Calibration: Develop methods for calibrating with and reconciling differences between LUH2, HILDA+, and other datasets to enable more consistent and validated outputs across time.

- Uncertainty Quantification: Implement a framework to assess and propagate uncertainty introduced at each processing step, improving interpretability and informing users of the confidence level in the simulated maps.

By addressing these challenges, the method can be further refined to provide a more accurate and comprehensive tool for studying historical land use and land cover dynamics on a global scale.

4.2 LAI

- **What has been achieved?**

We have demonstrated that statistical methods can offer a feasible approach for historical LAI reconstruction using existing input data. Despite the model we developed lacking a sophisticated architecture, extensive hyperparameter tuning, and key climate inputs—such as precipitation, temperature, and solar radiation—as well as geomechanical information like elevation, morphology, soil type, and texture, it still manages to capture many key aspects of the original dataset. As a result, we have successfully generated an extension of the CONFESS LAI dataset, pushing its historical coverage back to 1925. Additionally, the model functions effectively as an emulator, capable of producing on-demand LAI values using only land-use (LU) data as input. This represents a step toward the tighter integration of machine learning methods with numerical approaches within the context of NWP (Numerical Weather Prediction) and ESM (Earth System Modeling) frameworks.

- **What are the shortcomings ?**

The main shortcomings of this methodology, and consequently the generated dataset, are as follows:

- Lack of Monthly Climate Inputs: The model is limited in its ability to respond to inter-annual changes due to the reliance on learned climatology and the annual nature of LU's data. The absence of monthly inputs such as key climate variables—precipitation, temperature, and solar radiation—prevents the model from capturing interannual variability. At best, the model can replicate the climatology, but it lacks the capacity to reflect year-to-year fluctuations.
- Tendency to Fit Regional Averages: The model's inclination to fit predictions toward the regional average, even with a good overall fit, results in the suppression of extreme high or low values. This significantly affects areas where such extremes are ecologically important. The issue is further exacerbated by the severe imbalance in the dataset, where 92% of the data points fall within the 0–1 LAI range.

- Independent Regional Training and Inference: Since the models are trained and inferred independently for fixed regions back to 1925, uncertainty grows over time, leading to divergence between neighboring regions. As inference extends further back in time, visible borders begin to emerge between these regions, which can have a strong impact—especially when analyzing long-term historical trends.

- **Future work**

For future work, we aim to address the shortcomings mentioned above.

First and foremost, we plan to incorporate monthly climate data alongside static geomechanical information. Providing both high-resolution dynamic and static inputs will help guide the models toward more realistic and dynamic predictions.

To tackle the border issue, an intermediate solution would be to use overlapping tiles to address the discontinuities. As for a more general solution, we propose developing a single global model or employing a moving-window approach, where a single model progressively shifts across the globe. This strategy allows us to utilize large volumes of data without encountering memory constraints, while gradually synchronizing LAI predictions across regions at each time step—minimizing the emergence of artificial borders as early as possible.

These tasks are already underway as a continuation of this project in collaboration with vla CONCERTO and TerraDT Horizon Europe projects.

4.3 Lake cover

- **What has been achieved?**

Decadal monthly maps based on monthly 30 m resolution input data were generated. The used input data is open, up to date, consistent in time, and very high resolution (15 to 100 m resolution). The developed methodology is automated, reliable and adaptable. In general, the generated maps have a good correlation with high horizontal resolution yearly datasets, i.e. ESA CCI (300 m), Copernicus CGLS (100 m), ESA WorldCover (10 m).

- **What are the shortcomings?**

Water input data use was challenging, which led to a more complex methodology than initially expected, i.e. missing data over land and over far away ocean; data limited from 78°N to 60°S; missing islands; unreliable data far north (e.g. Greenland). All these issues were successfully overcome, yet some minor errors might still be present and there is a need for alternative reliable information for comparison and further validation.

Inland water is separated by a static mask adapted from Copernicus GLO30 representing 2015, which leads to constant ocean borders, e.g. new islands and/or coastal line erosion interchange with inland water. This suggests the need for a better computationally cheap ocean and inland water separation.

- **Future work**

Currently, long run (from February 1939 til December 2019) offline with and without data assimilation experiments are in progress:

- 'control' with static lake and land sea covers, static vegetation cover and type, recent time climatological leaf area index (LAI);
- 'vegetation' with static lake and land sea covers, yearly varying vegetation cover and type, yearly varying LAI based on land cover land use changes;

- 'lake' with time-varying lake and land sea covers, static vegetation cover and type, recent time climatological leaf area index (LAI);
- 'all' with time-varying lake and land sea covers, yearly varying vegetation cover and type, yearly varying LAI based on land cover land use changes.

We plan to analyze these results and compare globally and regionally with available high fidelity satellite observations.

References

Abernathey, R. P., Augspurger, T., Banihirwe, A., Busecke, J. J. M., Camargo, P., Cazenave et al. (2021). Cloud-Native Repositories for Big Scientific Data. *Computing in Science & Engineering*, 23(2), 85–95. https://doi.org/10.1109/MCSE.2021.3059437

Alessandri, A., Catalano, F., De Felice, M., et al. (2017). Multi-scale enhancement of climate prediction over land by increasing the model sensitivity to vegetation variability in EC-Earth. *Climate Dynamics*, 49, 315–1237.

Boussetta, S., Balsamo, G., Beljaars, A., et al. (2015). Assimilation of surface albedo and vegetation states from satellite observations and their impact on numerical weather prediction. *Remote Sensing of Environment*, 163, 111–126.

Boussetta S., Balsamo G. (2023). Leaf Area Index data for the period of 1993 to 2019 based on harmonization of the CGLS/C3S data and the AVHRR based data [Dataset]. https://confess-h2020.eu/results/data-sets/

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Carrea, L., Crétaux, J.-F., Liu, X., Wu, Y., Bergé-Nguyen, M., Calmettes, B., … & Zhang, D. (2024). ESA Lakes Climate Change Initiative (Lakes_cci): Lake products, Version 2.1 [Dataset]. NERC EDS Centre for Environmental Data Analysis. https://dx.doi.org/10.5285/7fc9df8070d34cacab8092e45ef276f1

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). https://doi.org/10.1145/2939672.2939785

Chini, L.P., G.C. Hurtt, R. Sahajpal, S. Frolking, K.K Goldewijk, S. Sitch, J. Pongratz, B. Poulter, L. Ma, and L. Ott. 2021. LUH2-GCB2019: Land-Use Harmonization 2 Update for the Global Carbon Budget, 850-2019. ORNL DAAC, Oak Ridge, Tennessee, USA. https://doi.org/10.3334/ORNLDAAC/1851

Cho, K., van Merriënboer, B., Gulcehre, C., et al. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint*, arXiv:1406.1078.

Duveiller, G., Baret, F., Cescatti, A., et al. (2022). Getting the leaves right matters for estimating temperature extremes. *Geoscientific Model Development Discussions*, 2022, 1–26.

Eerola, K., Rontu, L., Kourzeneva, E., Kheyrollah Pour, H., & Duguay, C. (2014). Impact of partly ice-free Lake Ladoga on temperature and cloudiness in an anticyclonic winter situation – A case study using a limited area model. *Tellus A: Dynamic Meteorology and Oceanography*, 66(1), 23929. https://doi.org/10.3402/tellusa.v66.23929

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, 14(2), 179–211.

Fang, H., Jiang, C., Li, W., Wei, S., Baret, F., & Myneni, R. B. (2019). An overview of global leaf area index (LAI): Methods, products, validation, and applications. *Reviews of Geophysics*, 57(3), 739–799.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780.

Hurtt, G. C., Chini, L., Sahajpal, R., et al. (2020). Harmonization of global land use change and management for the period 850–2100 (LUH2) for CMIP6. *Geoscientific Model Development*, 13, 5425–5464. https://doi.org/10.5194/gmd-13-5425-2020

Kimpson, T., Choulga, M., Chantry, M., Balsamo, G., Boussetta, S., Dueben, P., & Palmer, T. (2023). Deep learning for quality control of surface physiographic fields using satellite Earth observations. *Hydrology and Earth System Sciences*, 27, 4661–4685. https://doi.org/10.5194/hess-27-4661-2023

Li, W., MacBean, N., Ciais, P., Defourny, P., Lamarche, C., Bontemps, S., Houghton, R. A., and Peng, S. (2018). Gross and net land cover changes in the main plant functional types derived from the annual ESA CCI land cover maps (1992–2015), Earth Syst. Sci. Data, 10, 219–234, https://doi.org/10.5194/essd-10-219-2018

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

Pekel, J.-F., Cottam, A., Gorelick, N., & Belward, A. (2016). High-resolution mapping of global surface water and its long-term changes. *Nature*, 540, 418–422. https://doi.org/10.1038/nature20584

Rouault, E., Warmerdam, F., Schwehr, K., Kiselev, A., Butler, H., Łoskot, M., Szekeres, T., Tourigny, E., Landa, M., Miara, I., Elliston, B., Chaitanya, K., Plesea, L., Morissette, D., Jolma, A., Dawson, N., Baston, D., de Stigter, C., & Miura, H. (2025). GDAL (v3.11.0). Zenodo. https://doi.org/10.5281/zenodo.15375292

Samuelsson, P., Kourzeneva, E., & Mironov, D. (2010). The impact of lakes on the European climate as simulated by a regional climate model. *Boreal Environment Research*, 15, 113–129. https://www.borenv.net/BER/archive/pdfs/ber15/ber15-113.pdf

Winkler, K., Fuchs, R., Rounsevell, M., Herold, M. (2020). HILDA+ Global Land Use Change between 1960 and 2019 [Dataset]. PANGAEA. https://doi.org/10.1594/PANGAEA.921846

Winkler, K., Fuchs, R., Rounsevell, M. et al. (2021). Global land use changes are four times greater than previously estimated. *Nature Communications*, 12, 2501. https://doi.org/10.1038/s41467-021-22702-2

Zarr Development Team. (2022). *Zarr Specification Version 2*. https://zarr.readthedocs.io/en/stable/spec/v2.html

Document History

Version	Author(s)	Date	Changes
0.1	Etienne Tourigny, Amirpasha Mozaffari, Vinayak Huggannavar, Iria Ayan, ,Margarita Choulga, Souhail Boussetta, David Fairbairn	May 2025	Initial version for internal review
1.0	As above	May 2025	Issued after internal review comments

Internal Review History

Internal Reviewers	Date	Comments
Patricia de Rosnay (ECMWF) Sofia Ermida (IPMA)	May 2025	Initial version