Historical Reconstruction and Future Projection of Land Surface Boundary Conditions

Amirpasha Mozaffari
Barcelona Supercomputing Center
Barcelona, Spain
amirpasha.mozaffari@bsc.es

Marina Castaño
Barcelona Supercomputing Center
Barcelona, Spain
marina.castano@bsc.es

Stefano Materia
Barcelona Supercomputing Center
Barcelona, Spain
stefano.materia@bsc.es

Amanda Duarte
Barcelona Supercomputing Center
Barcelona, Spain
amanda.duarte@bsc.es

Abstract

Uncertainty in the terrestrial carbon cycle remains a major constraint in climate projections, partly driven by the uncertainties affecting the land surface representation and variability in Earth system models. To address this limitation, we propose a data-driven framework for generating high-resolution historical reconstructions and future projections of key land surface variables. The framework will follow a two-phase approach using a U-Net architecture. In the first phase, it will reconstruct annual land use and land cover by integrating coarse-resolution scenario data and climate reanalysis with static geophysical features. In the second phase, the resulting high-resolution maps will be used to predict dynamic biophysical variables, particularly leaf area index, at finer temporal scales. Trained on Earth Observation data, the models learn to reproduce spatially explicit and physically consistent land surface patterns, extending temporal coverage to periods lacking direct observations. The final product will be a suite of open-source emulators designed for real-time coupling with digital twin platforms, such as those developed under the Destination Earth initiative. By delivering realistic and evolving land surface conditions on demand, this work aims to reduce critical uncertainties and improve the predictive power of next-generation climate simulations.

1 Introduction

The terrestrial carbon cycle remains a major source of uncertainty in climate projections [3], partly because land surface processes are not fully resolved. Land-use (LU) and land-cover (LC) changes are significant sources of uncertainty in estimating carbon fluxes [10]. Relying on coarse or outdated land boundary data can further lead to misrepresentation of land–atmosphere exchanges. Studies have shown that coarse land cover spatial resolutions can introduce substantial biases in simulated terrestrial carbon sequestration, affecting its magnitude, interannual variability, and spatial distribution [33]. By contrast, providing high-resolution ($\approx$1 km) land surface information, such as detailed land cover maps, leaf area index (LAI), or satellite vegetation indices like Normalized Difference Vegetation Index (NDVI), enhanced vegetation index (EVI), and soil-adjusted vegetation index (SAVI), allows climate models to capture fine-scale heterogeneity in vegetation and soil characteristics. This improved detail enhances the realism of carbon, water, and energy fluxes between the land and atmosphere. For example, using a new 1 km-resolution land parameter dataset in a land model produced pronounced spatial variability in soil moisture and surface fluxes (latent heat and radiation),

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2025.

whereas aggregating those inputs to ∼12 km led to a loss of about 31–54% of the spatial information [17]. High-resolution vegetation data are particularly crucial, since changes in LAI or greenness directly affect processes like photosynthesis (carbon uptake), evapotranspiration, and surface energy balance. For instance, a drop in LAI reduces canopy shading, increases ground net radiation, and dries out soil moisture [4, 9]. Accurate representation of high-resolution land boundary conditions in climate models is essential for reducing uncertainties in the terrestrial carbon cycle and for improving simulations of coupled carbon–water–energy fluxes [10]. Complementing this, advances in remote sensing of vegetation provide critical observational constraints that enhance the monitoring, understanding, and modelling of these land–atmosphere interactions [26]. Several datasets offer valuable insights into land surface boundaries, but each falls short in terms of spatial resolution, temporal (historical or future), and global coverage, or continuity [13, 7, 30, 19, 20, 5]. Building on these insights and datasets, our work aims to reduce uncertainties in terrestrial carbon cycle representation by developing high-resolution land surface boundary datasets that span historical, contemporary, and future periods. By integrating satellite-era observations with reconstructions for pre-observation times and projections for observation-limited futures, we provide temporally continuous, spatially detailed datasets for use in climate and weather models. These datasets are designed to improve the representation of land surface heterogeneity, enabling more accurate simulations of carbon, water, and energy fluxes over time.

2 Related Approaches

A broad spectrum of machine learning (ML) techniques is increasingly employed for historical reconstruction, gap filling, and future projection of land surface variables such as LU, LC, LAI, NDVI, EVI, and SAVI. These approaches have expanded the capacity to generate temporally and spatially continuous datasets, particularly in regions with sparse observations. Among them, tree-based methods—particularly random forests (RF), have been widely adopted for diverse tasks including cropland reconstruction [31], LU/LC change estimation [1], high-resolution NDVI reconstruction [28], and the downscaling and gap filling of vegetation datasets [29]. RF has also been utilized to model vegetation health responses under climate variability [15]. XGBoost, another ensemble-based method, has shown strong performance in the historical reconstruction of LC and LAI [22]. Meanwhile, recurrent architectures such as long short-term memory (LSTM) networks have demonstrated effectiveness in reconstructing historical time series [32], forecasting LU change [34], and estimating LAI [21, 18]. In parallel, convolutional neural networks (CNNs) have been employed for spatial downscaling of remote sensing products [14] and regional-scale prediction of LAI [16], underscoring their ability to capture spatial hierarchies and fine-scale variability.

3 Proposed Approaches

To generate temporally continuous, spatially detailed land surface boundaries, we propose a two-step approach: first, reconstruct slow-varying variables (LU and LC), then use these outputs to reconstruct and predict fast-varying components such as LAI.

3.1 Phase One: High-Resolution LU/LC Prediction

In the first phase, we aim to generate historical and future annual LU and LC maps at high spatial resolution (1 km), derived from coarse-resolution inputs (0.25°, ∼28 km). The goal is to produce high-resolution LU maps for a target year t (H_t), with each pixel classified into predefined categories (e.g., Forest, Cropland, Urban). Ground-truth labels are sourced from the high-resolution HILDA+ LU/LC dataset [30]. The input tensor for year t is a multichannel stack comprising Land-Use Harmonization 2 (LUH2) data for year t [13], static topographic features (elevation and bathymetry) [11], and an autoregressive prior from year $t-1$ or $t+1$ to enable bidirectional prediction. We implemented an initial version of the LU using a U-Net architecture [27] for 512-by-512-pixel samples and discussed it in more detail in section A. Additional features will include current and projected Köppen-Geiger climate classifications [2], as well as static soil characteristics such as texture and type. For annual reconstruction and projection of high-resolution LC, we will use the upscaled ESA CCI LC dataset [6] at 1 km resolution as ground truth. Additional predictors include climate variables—annual temperature, precipitation, solar radiation, and soil moisture—sourced from ERA5 [12] and ERA5 Land [23].

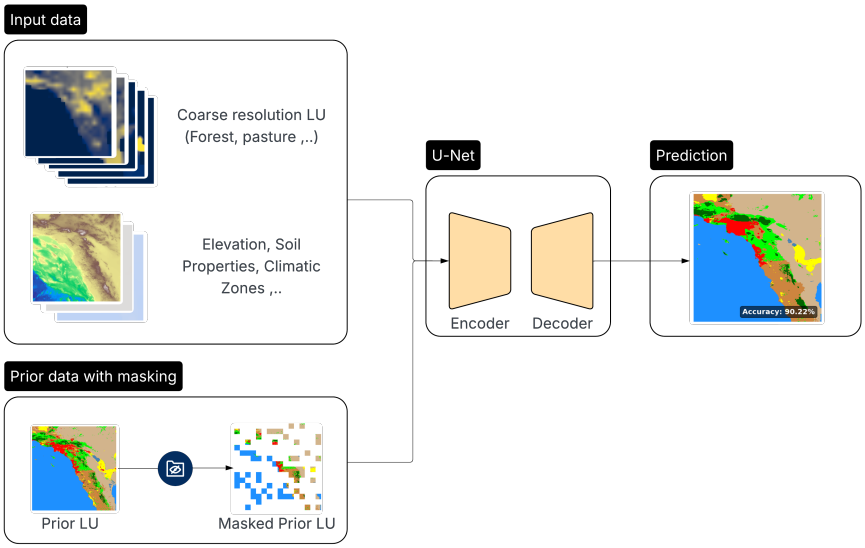

Figure 1: Overview of the LU reconstruction pipeline: A U-Net model takes as input coarse-resolution LUH2 land-use classes (e.g., forest, pasture), land surface parameters, high-resolution climatic zones, and static environmental variables (e.g., elevation, slope, soil). Masked land-use data from an adjacent year, preceding for reconstruction or following for forecasting, is added as auxiliary input. The standard U-Net then segments and predicts high-resolution land-use classes for the target year.

3.2 Phase Two: High-Resolution Dynamic Biophysical Parameter Prediction

Building on the results of phase one, this stage extends the framework to predict continuous biophysical variables characterizing vegetation state. Specifically, we aim to generate high-resolution maps of leaf area index (LAI_t) for a given time step t (monthly or sub-monthly). Ground-truth data are provided by the ESA CCI LAI product [7, 8]. The input tensor includes all features from phase one, augmented with high-frequency climate data, monthly or sub-monthly records from ERA5 [12] and ERA5 Land [23], and atmospheric CO_2 concentrations as [25], as well as a prior mask of LAI. These variables directly influence vegetation dynamics and phenology. While the core U-Net architecture is retained, we also explore sequential models to capture the strong temporal dependencies inherent in vegetation processes. We investigate the feasibility and limitations of a two-phase training strategy. Specifically, we compare separate LAI models, each using the outputs of the phase one model as input, with a joint multi-task model that simultaneously predicts LU, LC, and LAI. We will evaluate both approaches in terms of accuracy, computational efficiency, error propagation, and uncertainty.

4 Pathway to Impact

In this study, we propose a methodology to estimate historical and future land surface boundary conditions essential for improving weather and climate prediction. Advancing data-driven modeling, these datasets will enable more accurate and reliable forecasting across the climate science community and support ultra-high-resolution digital twins of the Earth. Within this framework, our models act as dynamic, online boundary condition generators, coupled in real time with digital twin infrastructure, producing high-resolution, time-evolving LU, LC, and LAI maps that respond to climate change. To ensure transparency, reproducibility, and broad engagement, all datasets and models will be released as open-source, including code and pretrained weights, supporting continued innovation in Earth system modeling.

Acknowledgments and Disclosure of Funding

This work is funded by Grant JDC2023-051208-I, funded by MICIU/AEI/10.13039/501100011033 and, as appropriate, by “ERDF A way of making Europe”, by “ERDF/EU”, by the “European Union”,

or by the "European Union NextGenerationEU/ PRTR". This work received funding from the European Union's Horizon Europe Framework Programme through the project CONCERTO (Grant Agreement 101185000). This work, as part of TerraDT - Digital Twin of Earth System for Cryosphere, Land Surface and related interactions, has received funding from the European Union's Horizon Europe Framework Programme (HORIZON) under Grant Agreement no. 101187992. "AD and SM acknowledge their AI4S fellowship within the "Generación D" initiative by Red.es, Ministerio para la Transformación Digital y de la Función Pública, for talent attraction (C005/24-ED CV1), funded by NextGenerationEU through PRTR".

References

[1] Bright Aboh and Alphonse Mutabazi. Satellite imagery analysis for land use, land use change and forestry: A pilot study in kigali, rwanda. In *NeurIPS 2020 Workshop on Tackling Climate Change with Machine Learning*, 2020. URL `https://www.climatechange.ai/papers/neurips2020/47`.

[2] Hylke E Beck, Niklaus E Zimmermann, Tim R McVicar, Noemi Vergopolan, Alexis Berg, and Eric F Wood. Present and future köppen-geiger climate classification maps at 1-km resolution. *Scientific data*, 5(1):1–12, 2018.

[3] Ben BB Booth, Chris D Jones, Mat Collins, Ian J Totterdell, Peter M Cox, Stephen Sitch, Chris Huntingford, Richard A Betts, Glen R Harris, and Jon Lloyd. High sensitivity of future global warming to land carbon cycle processes. *Environmental Research Letters*, 7(2):024002, 2012.

[4] Souhail Boussetta, Gianpaolo Balsamo, Anton Beljaars, Tomas Kral, and Lionel Jarlan. Impact of a satellite-derived leaf area index monthly climatology in a global numerical weather prediction model. *International journal of remote sensing*, 34(9-10):3520–3542, 2013.

[5] Min Chen, Chris R Vernon, Neal T Graham, Mohamad Hejazi, Maoyi Huang, Yanyan Cheng, and Katherine Calvin. Global land use for 2015–2100 at 0.05 resolution under diverse socioeconomic and climate scenarios. *Scientific Data*, 7(1):320, 2020.

[6] Copernicus Climate Change Service. Land cover classification gridded maps from 1992 to present derived from satellite observation. `https://doi.org/10.24381/cds.006f2c9a`, 2019. Accessed on 26-Jul-2025.

[7] Copernicus Global Land Service / EEA. Leaf Area Index 1999–2020 (raster 1 km), global, 10-daily – version 2, 2020. URL `https://land.copernicus.eu/en/products/vegetation/leaf-area-index-v2-0-1km`. Temporal coverage: 1999–2020; spatial resolution: raster 1 km; dekadal (every 10 days).

[8] Copernicus Global Land Service / European Commission's Joint Research Centre. Leaf Area Index 2014–present (raster 300 m), global, 10-daily – version 1, 2023. URL `https://land.copernicus.eu/en/products/vegetation/leaf-area-index-300m-v1.0`. Temporal coverage: 2014–present; spatial resolution: raster 300 m; dekadal (every 10 days); regularly updated.

[9] Hongliang Fang, Frederic Baret, Stephen Plummer, and Gabriela Schaepman-Strub. An overview of global leaf area index (lai): Methods, products, validation, and applications. *Reviews of Geophysics*, 57(3):739–799, 2019.

[10] Bettina K Gier, Manuel Schlund, Pierre Friedlingstein, Chris D Jones, Colin Jones, Sönke Zaehle, and Veronika Eyring. Representation of the terrestrial carbon cycle in cmip6. *Biogeosciences*, 21(22):5321–5360, 2024.

[11] GEBCO Compilation Group. Gebco 2024 grid. Distributed by GEBCO, British Oceanographic Data Centre, 2024. URL `https://www.gebco.net/data-products/gridded-bathymetry-data`. Accessed on 27 July 2025.

[12] Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The era5 global reanalysis. *Quarterly journal of the royal meteorological society*, 146(730):1999–2049, 2020.

[13] George C Hurtt, Louise Chini, Ritvik Sahajpal, Steve Frolking, Benjamin L Bodirsky, Katherine Calvin, Jonathan C Doelman, Justin Fisk, Shinichiro Fujimori, Kees Klein Goldewijk, et al. Harmonization of global land-use change and management for the period 850–2100 (luh2) for cmip6. *Geoscientific Model Development Discussions*, 2020:1–65, 2020.

[14] Yuanxin Jia, Yong Ge, Yuehong Chen, Sanping Li, Gerard BM Heuvelink, and Feng Ling. Super-resolution land cover mapping based on the convolutional neural network. *Remote Sensing*, 11(15):1815, 2019.

[15] Fithrothul Khikmah, Christoph Sebald, Martin Metzner, and Volker Schwieger. Modelling vegetation health and its relation to climate conditions using copernicus data in the city of constance. *Remote Sensing*, 16(4):691, 2024.

[16] Hao Li, Yuyu Zhou, Xiang Zhao, Xin Zhang, and Shunlin Liang. A dataset of 0.05-degree leaf area index in china during 1983–2100 based on deep learning network. *Scientific Data*, 11(1): 1122, 2024.

[17] Lingcheng Li, Gautam Bisht, Dalei Hao, and L Ruby Leung. Global 1 km land surface parameters for kilometer-scale earth system modeling. *Earth System Science Data*, 16(4): 2007–2032, 2024.

[18] Tian Liu, Huaan Jin, Ainong Li, Hongliang Fang, Dandan Wei, Xinyao Xie, and Xi Nan. Estimation of vegetation leaf-area-index dynamics from multiple satellite products through deep-learning method. *Remote sensing*, 14(19):4733, 2022.

[19] Meng Luo, Guohua Hu, Guangzhao Chen, Xiaojuan Liu, Haiyan Hou, and Xia Li. 1 km land use/land cover change of china under comprehensive socioeconomic and climate scenarios for 2020–2100. *Scientific data*, 9(1):110, 2022.

[20] Jiaying Lv, Yifan Gao, Changqing Song, Li Chen, Sijing Ye, and Peichao Gao. Land system changes of terrestrial tipping elements on earth under global climate pledges: 2000–2100. *Scientific Data*, 12(1):163, 2025.

[21] Han Ma and Shunlin Liang. Development of the glass 250-m leaf area index product (version 6) from modis data using the bidirectional lstm deep learning model. *Remote sensing of environment*, 273:112985, 2022.

[22] Amirpasha Mozaffari, Stefano Materia, Vinayak Huggannavar, Lina Teckentrup, Iria Ayan, Etienne Tourigny, and Markus Donat. Reconstruction and downscaling of historical land surface boundary conditions with machine learning. Technical report, Copernicus Meetings, 2025.

[23] Joaquín Muñoz-Sabater, Emanuel Dutra, Anna Agustí-Panareda, Clément Albergel, Gabriele Arduini, Gianpaolo Balsamo, Souhail Boussetta, Margarita Choulga, Shaun Harrigan, Hans Hersbach, et al. Era5-land: A state-of-the-art global reanalysis dataset for land applications. *Earth system science data*, 13(9):4349–4383, 2021.

[24] D. J. Newman. Zarr storage specification version 2: Cloud-optimized persistence using zarr. Technical report, NASA Earth Science Data and Information System Standards Coordination Office, 2024. URL `https://doi.org/10.5067/DOC/ESCO/ESDS-RFC-048v1`.

[25] NOAA Global Monitoring Laboratory. Monthly average mauna loa co_2: June 2025 = 429.61 ppm. `https://gml.noaa.gov/ccgg/trends/`, July 2025. URL `https://gml.noaa.gov/ccgg/trends/`. Accessed on 27 July 2025.

[26] I Colin Prentice, Manuela Balzarolo, Keith J Bloomfield, Jing M Chen, Benjamin Dechant, Darren Ghent, Ivan A Janssens, Xiangzhong Luo, Catherine Morfopoulos, Youngryel Ryu, et al. Principles for satellite monitoring of vegetation carbon uptake. *Nature Reviews Earth & Environment*, 5(11):818–832, 2024.

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241. Springer, 2015.

[28] Mengmeng Sun, Adu Gong, Xiang Zhao, Naijing Liu, Longping Si, and Siqing Zhao. Reconstruction of a monthly 1 km ndvi time series product in china using random forest methodology. *Remote Sensing*, 15(13):3353, 2023.

[29] Barry Van Jaarsveld, Sandra M Hauswirth, and Niko Wanders. Machine learning and global vegetation: random forests for downscaling and gap filling. *Hydrology and Earth System Sciences*, 28(11):2357–2374, 2024.

[30] Karina Winkler, Richard Fuchs, M Rounsevell, and Martin Herold. Hilda+ global land use change between 1960 and 2019 [dataset]. pangaea, 2020.

[31] Jia Yang, Bo Tao, Hao Shi, Ying Ouyang, Shufen Pan, Wei Ren, and Chaoqun Lu. Integration of remote sensing, county-level census, and machine learning for century-long regional cropland distribution data reconstruction. *International Journal of Applied Earth Observation and Geoinformation*, 91:102151, 2020.

[32] Jie Yang and Xin Huang. 30 m annual land cover and its dynamics in china from 1990 to 2019. *Earth System Science Data Discussions*, 2021:1–29, 2021.

[33] SQ Zhao, S Liu, Z Li, and Terry L Sohl. A spatial resolution threshold of land cover in estimating terrestrial carbon sequestration in four counties in georgia and alabama, usa. *Biogeosciences*, 7 (1):71–80, 2010.

[34] Xin Zhao, Ping Wang, Songhe Gao, Muhammad Yasir, and Qamar Ul Islam. Combining lstm and plus models to predict future urban land use and land cover change: A case in dongying city, china. *Remote Sensing*, 15(9):2370, 2023.

A Appendix

Preliminary Results

In phase one, we developed the backbone for LU reconstruction and projection using a U-Net architecture. A preprocessing pipeline was established to align all input datasets to the grid and projection of the target dataset HILDA+ [30], at 1 km resolution in WGS84. All datasets were stored as 3D cubes in Zarr format [24], chunked in 512×512-pixel blocks for efficient retrieval. Initial inputs included coarse-resolution LUH2h data comprising 12 fractional LU variables (ranging from 0 to 1) and two land surface parameters standardized to unit variance on a $0.25^\circ \times 0.25^\circ$ ($\sim$31 km) grid [13]. High-resolution GEBCO data provided elevation and bathymetry [11]. The target HILDA+ dataset originally contained 13 classes, which we consolidated into 8 LU categories for this stage. A prior LU map, partially masked with randomly selected 32×32-pixel patches, was added as an auxiliary input channel.

The model employed a standard U-Net with 16 base channels, accepting 512×512-pixel inputs. Its encoder consisted of four max-pooling and double-convolution blocks, mirrored by a symmetric decoder with upsampling and skip connections, terminating in a 1×1 convolution for final segmentation. Trained on 30,000 samples, the model achieved promising performance. Figure 2 shows LU reconstruction for San Diego in 2001 using LUH2h, elevation, and masked 2000 LU inputs, achieving 90% accuracy under 75% masking. In addition to pixel-wise accuracy, performance was assessed using the mean Intersection over Union (mIoU), yielding a strong score of 0.626 for this case. This reflects robust performance on dominant landscape classes, though confusion persists between similar vegetation types (e.g., forest vs. grass/shrubland) and in detecting small features such as inland water bodies. This pattern was consistent across test regions, where the model achieved an average mIoU of 0.42, highlighting the persistent challenge of identifying underrepresented LU types.

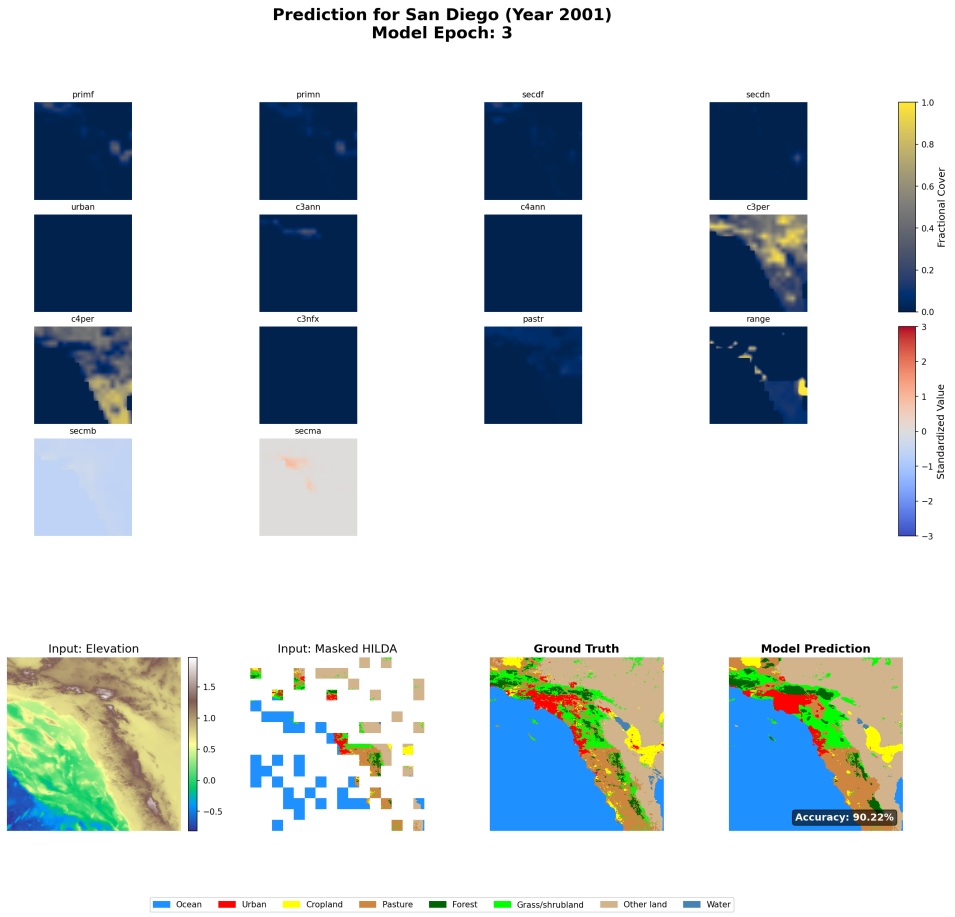

Figure 2: Reconstruction of the LU map for the city of San Diego for the year 2001, using LUH2h fractional LU with 25 by 25 resolution, an elevation map, and a prior LU map randomly 75 percent masked with 32 by 32 pixel patches.