CLIMATE ANALYSIS IN AFRICAN CITIES (CAIAC)

D2.2 Tailored Data Delivery Note

Study accomplished under the authority of the European Space Agency
Reference: 2025/EI/R/3452
March 2026

Vision on technology for a better world

vito.be

D2.2 Tailored Data Delivery Note

VITO
Boeretang 200
2400 MOL
Belgium
VAT No: BE0244.195.916
vito@vito.be – www.vito.be
IBAN BE34 3751 1173 5490 BBRUBEBB

Nele Veldeman
Project Manager
nele.veldeman@vito.be

Koen De Ridder
Science Lead
koen.deridder@vito.be

AUTHORS

VITO

Benjamin Lanssens

Tomas Crols

Koen De Ridder

Jonathan Leon Tavares

Niels Souverijns

Inge Uljee

Nele Veldeman

University of Ottawa

Mohammad Alobaidi

Hossein Bonakdari

Ousmane Seidou

DISTRIBUTION LIST

Clement Albergel, ESA

Amy Campbell, ESA

SUMMARY

This deliverable (D2.2 - DAT1) documents the first tailored data package produced in CAIAC to support the urban heat (UrbClim), urban growth (GeoDynamiX) and urban flood modelling (SAHEL) chains. It describes how heterogeneous Earth Observation (EO) datasets, reanalysis fields and climate projections are harmonised onto a common city-specific modelling grid, quality-controlled, and delivered in model-ready formats for UrbClim, GeoDynamiX and SAHEL.

DAT1 provides, for each target city, the following core artefacts:

- A reproducible definition of the modelling grid (projection, grid spacing and domain bounds), together with a city grid index and optional domain polygons.
- EO-derived terrain and land-surface parameters required by UrbClim (surface NetCDF) including land use, vegetation cover, terrain height, impervious fraction, aerodynamic roughness, anthropogenic heat flux, and soil hydraulic parameters.
- Present-day meteorological forcing for UrbClim derived from ERA5; and a future-forcing preparation workflow based on bias-corrected CMIP6 deltas (delta matrices and supporting artefacts).
- GeoDynamiX calibration inputs (land-cover 'snapshot' of historical years) and scenario outputs (annual land-use and population maps).
- Flood-modelling support layers and precipitation-extremes products (return-period rainfall estimates) required by the SAHEL flood-modelling workflow, together with supporting quality-control information.

TABLE OF CONTENTS

LIST OF ACRONYMS

Acronym	Definition
AHF	Anthropogenic Heat Flux
AOI	Area of Interest
CAIAC	Climate Analysis in African Cities
CCI	Climate Change Initiative
CMA-ES	Covariance Matrix Adaptation Evolution Strategy
CMIP6	Coupled Model Intercomparison Project Phase 6
CRS	Coordinate Reference System
DAT1	Data Tailoring package 1
DEM	Digital Elevation Model
ECMWF	European Centre for Medium-Range Weather Forecasts
ECV	Essential Climate Variable
EO	Earth Observation
EPSG	Coordinate reference system identifier (EPSG code)
ERA5	ECMWF Reanalysis version 5
GDX	GeoDynamiX (urban growth) modelling chain
GEV	Generalised Extreme Value distribution
GHSL	Global Human Settlement Layer
GISA	Global Impervious Surface Area
GSHHS	Global Self-consistent Hierarchical High-resolution Shorelines
HPC	High Performance Computing
IDF	Intensity-Duration-Frequency
KO	Kick-off
LCZ	Local Climate Zone
LST	Land Surface Temperature
NDVI	Normalised Difference Vegetation Index
NetCDF	Network Common Data Form
OSM	OpenStreetMap
QA	Quality Assurance
QC	Quality Control
RMSE	Root Mean Square Error
SAHEL	Satellite-based African Hydrological Ensemble Learner
SLURM	Simple Linux Utility for Resource Management
SSP	Shared Socioeconomic Pathway
SST	Sea Surface Temperature
UHI	Urban Heat Island
UrbClim	Urban boundary-layer climate model
UTM	Universal Transverse Mercator
WGS84	World Geodetic System 1984

1 INTRODUCTION

Climate Analysis in African Cities (CAIAC) develops a consistent modelling framework to assess present and future heat and flood risks across a set of African cities. The framework combines UrbClim for urban climate simulations, GeoDynamiX for urban growth projections and SAHEL for flood hazard mapping and forecasting. Each model requires spatially and temporally harmonised input data derived from multiple Earth Observation (EO) products and climate datasets.

Task 2.2 ('Data tailoring') provides this harmonisation by defining a common city-specific modelling grid, reprojecting and regridding input datasets, applying quality control and filtering, and providing the resulting datasets in model-ready formats.

The present document focuses on the concrete datasets and processing workflows delivered under DAT1:

- Definition of the target modelling grids and domain selection rules used across all city products (Section 2).
- EO-derived terrain and land-surface inputs for UrbClim, including land-use harmonisation and derivation of key terrain parameters (Section 3.1).
- Meteorological forcing for UrbClim derived from ERA5 reanalysis and the preparation workflow for future forcing using bias-corrected CMIP6 deltas (Section 3.2).
- GeoDynamiX calibration inputs (land-cover snapshot of historical years) and annual scenario outputs (land use and population maps) supporting future exposure and feedback analyses (Section 3.3).
- Flood-model support datasets including precipitation-extremes/return-period products and static supporting layers (Section 3.4).

Dataset versions, access points and licensing are primarily maintained in the CAIAC Data Inventory (Deliverable D1.3). Where additional workflow-specific datasets are used but not fully described in D1.3, concise inventory-style descriptions are included directly in this deliverable in section 2.2. The scientific rationale and modelling strategy are described in the Scientific Methodology Strategy (Deliverable D2.1). This D2.2 note complements those deliverables by documenting the implementation details, quality assurance and delivered datasets.

2 DATA TAILORING WORKFLOW

Across all CAIAC cities, data tailoring follows a consistent sequence of steps:

1. selection of source datasets and configuration parameters.
2. definition of the target city grid and modelling domain.
3. regridding, resampling and conversion to model-specific formats.
4. quality control and filtering.
5. uncertainty characterisation.

The workflow is implemented through Python processing scripts, configuration files and batch jobs executed on the project processing infrastructure. Only the operational rules that affect the delivered artefacts are documented here.
Sections 2.1 to 2.6 summarise the workflow elements that are common across datasets. Section 3 then documents the dataset-specific tailoring chains, outputs, QC and uncertainty notes retained in this deliverable.

2.1 Source datasets

DAT1 integrates EO, ancillary geospatial layers, reanalysis fields and climate projections. The tailoring approach prioritises widely used, well-documented products (including ESA ECV datasets where applicable) and preserves, as far as possible, the native information content of each source while harmonising spatial reference, resolution and variable conventions.
The source datasets used in Task 2.2 are not re-described exhaustively where they are already documented in Deliverable D1.3. In those cases, the relevant D1.3 section is cited in-text when the dataset is first mentioned in the tailoring workflow. For datasets that are used in D2.2 but are not fully described as standalone inventory entries in D1.3, this deliverable provides additional inventory-style metadata, including source, spatial and temporal coverage, key variables, access/licensing information, and uncertainty considerations (section 2.2).
For the purposes of the present note, the relevant source groups are land-cover and built-up layers used in surface-parameter generation, vegetation and terrain products, ERA5 reanalysis used for present-day UrbClim forcing, bias-corrected CMIP6 data used for future perturbation artefacts, and hydroclimatic predictors used in the flood workflow. Dataset-specific implementation details are provided only in the corresponding subsections of Section 3.

2.2 Additional auxiliary datasets description

2.2.1 Global Impervious Surface Area (GISA-10m)

GISA-10m provides global impervious-surface information at 10 m resolution derived from Sentinel optical/radar imagery. In CAIAC, it is used to derive soil sealing / impervious fraction for UrbClim terrain files.

Core metadata

Source	Huang et al. (2022)
Spatial resolution	10 m
Native CRS	WGS 84 (EPSG: 4326)
Temporal coverage	2016
Update frequency	Versioned research dataset
Variables	Impervious surface / built-up impervious fraction or class

File format	GeoTIFF / raster tiles
Access / DOI	Zenodo, DOI: 10.5194/essd-14-3649-2022
Licence	CC-BY 4.0

Validation and uncertainty

GISA-10m was validated using independent test samples that were not used for training. The product reports an overall accuracy above 86%, with a global overall accuracy of 86.06% in the visually interpreted validation sample. The validation also included comparisons with existing global impervious-surface products and specific checks for rural, urban and arid-region performance. At continental scale, the average overall accuracy exceeded 85% for each continent. For different city-size classes, the reported overall accuracies were 85.35% for small cities, 87.43% for medium cities and 85.42% for large cities. In arid regions, which are particularly relevant for many African cities, GISA-10m showed improved detection of impervious surfaces compared with several existing products, although confusion between impervious surfaces, bare soil, compacted surfaces and roads may still occur.

2.2.2 Anthropogenic Heat Flux

An anthropogenic Heat Flux (AHF) raster is used in CAIAC as a static urban heat source term in UrbClim.

Core metadata

Source	Wang et al. (2022) global anthropogenic heat flux dataset
Spatial resolution	500 m
Native CRS	WGS 84 (EPSG: 4326)
Temporal coverage	2016
Update frequency	Versioned research dataset
Variables	Anthropogenic heat flux in W m^{-2}
File format	GeoTIFF / raster tiles
Access / DOI	National Tibetan Plateau/Third Pole Environment Data Center, DOI: 10.17616/R31NJMQT
Licence	CC-BY 4.0

Validation and uncertainty

AHF is model-derived and usually depends on population, energy use, transport and built-up proxies. Uncertainty is expected to be high in African cities where energy-use statistics, informal settlements and local activity patterns may not be well represented. In CAIAC, AHF is used as a static annual-mean field and should be interpreted as a first-order urban heat-source approximation rather than a temporally resolved estimate.

2.2.3 GSHHS / GSHHG Shoreline and Land-Sea Mask

The Global Self-consistent, Hierarchical, High-resolution Geography database (GSHHS/GSHHG) provides global shoreline, lake and island polygons. In CAIAC, it is used as an auxiliary land-sea mask to distinguish inland water from ocean water in coastal UrbClim domains.

Core metadata

Source	Wessel & Smith / SOEST University of Hawaiʻi / NOAA
Spatial resolution	Five vector resolutions: crude, low, intermediate, high and full

Native CRS	WGS 84 (EPSG: 4326)
Temporal coverage	Static coastline database
Update frequency	Versioned releases
Variables	Coastlines, lakes, islands in lakes, ponds in islands in lakes
File format	Binary/vector shoreline files
Access / DOI	NOAA / GSHHG distribution, DOI:10.1029/96JB00104
Licence	GNU Lesser General Public License

Validation and uncertainty
GSHHG is a compiled global shoreline database and does not provide a pixel-level uncertainty layer. It is suitable for broad land-sea masking at the UrbClim grid scale, but uncertainty may occur in estuaries, deltas, small islands, harbours, narrow channels and recently modified coastlines. In CAIAC, the resulting land-sea mask is therefore checked visually through the terrain-processing diagnostic plots.

2.2.4 GRDC River Discharge Data

The Global Runoff Data Centre (GRDC), hosted by the German Federal Institute of Hydrology (BfG), provides quality-controlled in-situ river discharge observations from national hydrological services. In CAIAC, GRDC discharge records are used as reference observations for calibrating, evaluating, or bias-adjusting river-discharge simulations and for checking the consistency of flood-related hydrological inputs.

Core metadata

Source	German Federal Institute of Hydrology
Spatial resolution	Station-based point observations
Native CRS	Station coordinates, typically geographic latitude/longitude
Temporal coverage	Variable by station; daily and/or monthly discharge records depending on station availability
Update frequency	Updated as new national hydrological-service data become available
Variables	River discharge, station metadata, catchment or subregion information where available
File format	Downloadable station time series, commonly text/CSV-style formats depending on portal export
Access / DOI	GRDC Data Portal
Licence	Research use; commercial use and redistribution are not allowed without permission

Validation and uncertainty
GRDC records are quality-checked in-situ discharge observations and are therefore suitable as hydrological reference data. The dataset does not provide a raster uncertainty layer. Uncertainty may arise from rating-curve changes, measurement gaps, station relocation, regulation or abstraction effects, and differences between the gauging location and the modelled urban catchment. Station coverage and record length vary strongly by basin and country.

2.2.5 ISIMIP3b Hydrological / River-Discharge Simulations

The Inter-Sectoral Impact Model Intercomparison Project phase 3b (ISIMIP3b) provides bias-adjusted climate forcing and impact-model outputs for historical and future climate-impact assessment. In the flood workflow, ISIMIP3b hydrological simulations can be used to provide river-discharge series under historical and future climate scenarios. These data support fluvial flood modelling where observed discharge records are incomplete or where future scenario-consistent discharge information is required.

Core metadata

Source	Inter-Sectoral Impact Model Intercomparison Project
Spatial resolution	Typically 0.5°
Native CRS	WGS 84 (EPSG: 4326)
Temporal coverage	Historical and future periods, depending on model and scenario
Update frequency	Versioned ISIMIP data releases
Variables	River discharge and related hydrological variables, depending on the selected hydrological model output
File format	NetCDF
Access / DOI	ISIMIP Repository, DOI: 10.48364/ISIMIP.230418.8
Licence	Depend on dataset, either CC BY 4.0 or CC0 1.0

Validation and uncertainty

ISIMIP3b river-discharge products are model outputs rather than observations. No standard pixel-level uncertainty layer is provided. Uncertainty should be characterised through the spread across climate models, scenarios, hydrological impact models and bias-adjustment assumptions.

2.2.6 NASA NEX-GDDP-CMIP6 / GDDP Climate Projections

NASA Earth Exchange Global Daily Downscaled Projections (NEX-GDDP) CMIP6 provides global, daily, bias-corrected and statistically downscaled CMIP6 climate projections. In CAIAC, the dataset supports the flood-modelling workflow by providing daily precipitation and temperature inputs for historical and future scenarios.

Core metadata

Source	NASA Earth Exchange / NASA Center for Climate Simulation
Spatial resolution	Approximately 0.25°
Native CRS	WGS 84 (EPSG: 4326)
Temporal coverage	Daily from 1950-01-01 to 2100-12-31
Update frequency	Versioned dataset releases
Variables	Includes precipitation, near-surface air temperature, minimum and maximum temperature, humidity, radiation and wind variables
File format	NetCDF
Access / DOI	AWS Open Data, DOI: 10.7917/OFSG3345
Licence	CC0 1.0

Validation and uncertainty

NEX-GDDP-CMIP6 is a bias-corrected and statistically downscaled climate-model product. It does not provide a standard pixel-level uncertainty layer. Uncertainty arises from the underlying Global Circulation Models, greenhouse-gas scenarios, bias-correction method, downscaling assumptions and observational reference data.

2.2.7 TAMSAT Rainfall Estimates

Tropical Applications of Meteorology using SATellite data and ground-based observations (TAMSAT) provides long-term rainfall estimates for Africa derived primarily from Meteosat thermal infrared observations calibrated against rain-gauge data. It is relevant for flood modelling because it offers spatially complete precipitation information over Africa, including regions where station observations are sparse.

Core metadata

Source	TAMSAT, University of Reading
Spatial resolution	~0.0375°
Native CRS	WGS 84 (EPSG: 4326)
Temporal coverage	1983 to near real time
Update frequency	Daily, pentadal, dekadal, monthly and seasonal products
Sensor lineage	Meteosat thermal infrared imagery calibrated with rain-gauge observations
Variables	Rainfall estimates, rainfall anomalies and related rainfall products depending on version
File format	NetCDF
Access / DOI	TAMSAT data service
Licence	CC BY 4.0

Validation and uncertainty

TAMSAT has been validated against rain-gauge observations and compared with other satellite rainfall products. The main product does not include an explicit pixel-level uncertainty layer. Uncertainty arises from the cloud-top-temperature rainfall-retrieval method, rain-gauge availability, calibration density, storm type and local orographic effects. The product performs best for convective rainfall systems and can underestimate warm-cloud or orographic precipitation. In CAIAC, TAMSAT is therefore used together with other precipitation and hydrological datasets where possible.

2.2.8 SRTM Digital Elevation Model

The Shuttle Radar Topography Mission (SRTM) provides near-global digital elevation data derived from radar interferometry. In CAIAC, SRTM is used as an auxiliary DEM in the flood workflow, particularly where Copernicus DEM coverage or suitability is limited.

Core metadata

Source	NASA / USGS / NGA / JPL-Caltech
Spatial resolution	30 m
Native CRS	WGS 84 (EPSG: 4326)
Temporal coverage	2000
Update frequency	Static dataset, with versioned processed releases
Sensor lineage	Shuttle Radar Topography Mission
Variables	Elevation above mean sea level
File format	GeoTIFF
Access / DOI	NASA Earthdata, DOI: 10.5067/MEASURES/SRTM/SRTMGL1.003
Licence	Openly shared, without restriction, in accordance with the EOSDIS Data Use and Citation Guidance.

Validation and uncertainty
SRTM provides consistent, near-global elevation information but is a radar-derived surface model and may include vegetation or built-structure effects. Errors may occur in steep terrain, urban areas, wetlands, water bodies and regions affected by radar shadow or void filling.

2.2.9 HWSD / SoilGrids Soil Properties

Soil properties are used in the flood workflow to support hydrological modelling, including infiltration, runoff generation and soil-water storage.

Core metadata

	SoilGrids	HWSD
Source	World Soil Information	FAO / IIASA / ISRIC / JRC and partners
Spatial resolution	250 m	Approximately 1 km
Native CRS	WGS 84 (EPSG: 4326)	WGS 84 (EPSG: 4326)
Temporal coverage	Static soil-property predictions	Static soil database
Update frequency	Versioned releases	Versioned releases
Variables	Sand, silt, clay, bulk density, organic carbon, pH, coarse fragments, CEC, nitrogen, uncertainty intervals	Soil mapping units and associated physical/chemical properties
File format	GeoTIFF	GeoTIFF
Access / DOI	SoilGrids portal	ISRIC
Licence	CC-BY 4.0	CC BY-NC-SA 3.0

Validation and uncertainty
SoilGrids is model-derived from global soil profile data and environmental covariates and includes uncertainty layers, commonly expressed through prediction intervals. HWSD is a harmonised legacy soil-unit database and may have coarser spatial detail.

2.3 Target grids

A list of city-specific CAIAC grids is adopted as the operational spatial reference for Task 2.2 outputs. The methodological rationale for domain selection is described in Deliverable D2.1; D2.2 records here only the final adopted rules and the resulting grid inventory used by the processing workflows.
For each city, the modelling grid is defined by the parameter set Longitude (clon), latitude (clat), spatial resolution (dxy), projection (CRS), with a default spatial resolution of 300 m. 600 m spatial resolution is required for selected very large cities by domain-size constraints.
Domain extent is selected to satisfy four criteria:

1. the present-day urban area is fully contained within the domain.
2. the domain includes a sufficient rural buffer to resolve the urban heat island signal.
3. projected urban expansion remains within the domain for the considered scenario and time horizon.
4. the maximum UrbClim domain size is constrained to 301 x 301 grid cells.

To minimise reprojection artefacts in the city core when ingesting ESA land-cover layers, the city-centre coordinate is adjusted ('snapped') to the nearest ESA land-cover pixel centre. The snapped coordinate centre is used to derive the local UTM zone and to define the grid geometry.

Table 1 Standard grid parameters and constraints

Parameter	Rule / value	Rationale / notes
Spatial resolution	300 m (default); 600 m (very large cities only)	Common resolution for harmonised EO inputs; optional coarser grid to extend domain size.
Number of pixels	Odd integer; constrained to [101, 301]	Odd n ensures a unique centre cell; n = 301 aligns with UrbClim domain-size constraint.
Projection	Local UTM EPSG for each city	Metric geometry for distances/areas; EPSG derived from city-centre longitude/latitude.
Centre coordinates	Snapping of city-centre to nearest ESA land-cover pixel centre	Reduces reprojection artefacts in the urban core when using ESA land-cover grids as a reference.

Exact grid geometry is computed as follows. Given $(clon, clat)$ and the city UTM projection (EPSG), the centre point is projected to (xc, yc) in metres. Let $h = (n - 1) / 2$. The projected bounds are then computed as $xmin = xc - h * dxy$, $xmax = xc + h * dxy$, $ymin = yc - h * dxy$, and $ymax = yc + h * dxy$. For polygon visualisation, bounds may be expanded by half a cell so that outlines correspond to outer cell edges.
Two special cases are implemented to ensure consistent domain handling:

1. Kinshasa and Brazzaville are treated as a coupled metropolitan area with a shared domain centre derived from the midpoint between their nominal centres.
2. resolution output of 600 m is used in Cairo, Luanda and Johannesburg where a 300 m domain would not provide sufficient rural buffer within the model domain-size constraint.

A city grid excel (city_grids.xlsx) is provided containing EPSG, dxy, n, domain width, centre lon/lat and projected bounds and an optional overview polygon product (KML/GeoJSON) for rapid quality control and communication. The city grid excel is reproduced in

Annex 2: City grid.

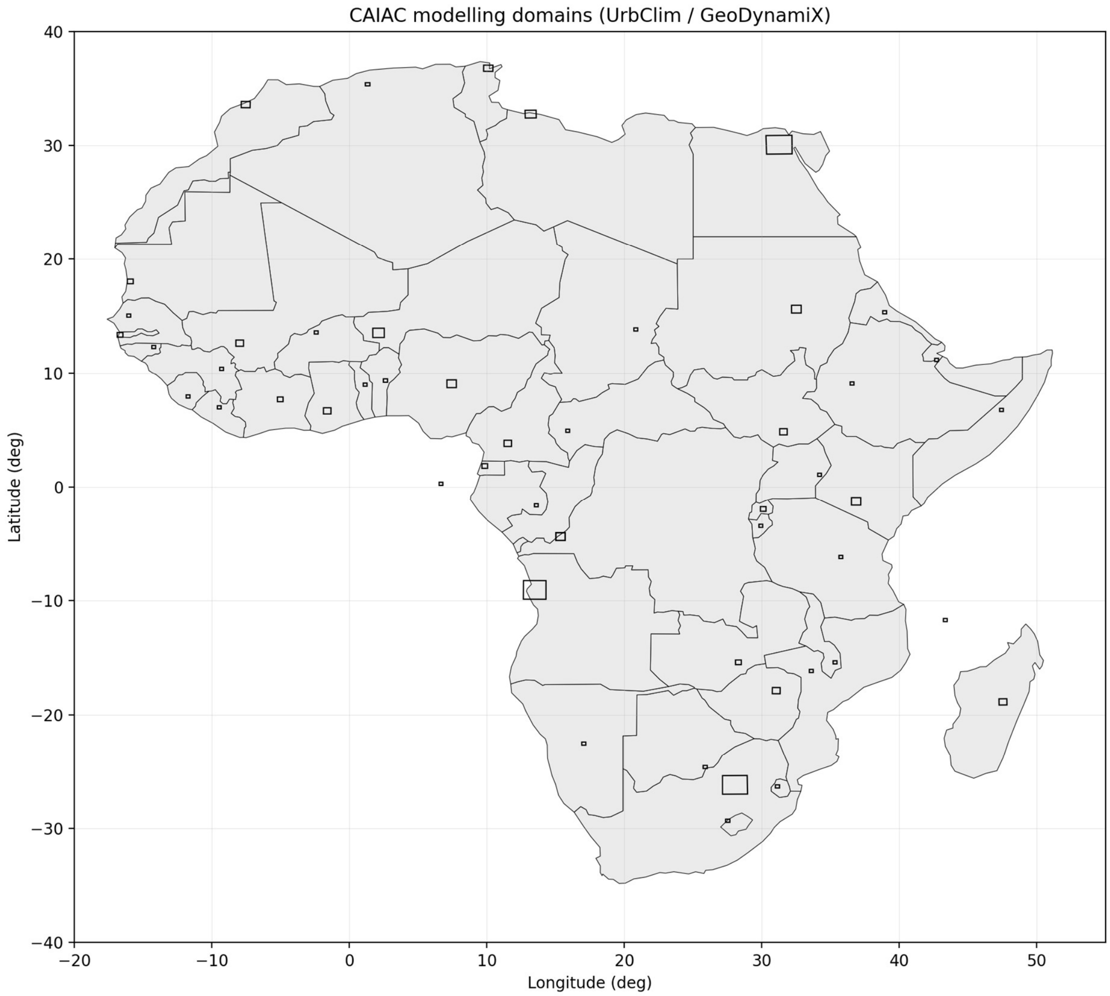


Figure 1 CAIAC 54 cities modelling grids.

2.4 Regridding and resampling methods

The detailed scientific rationale for grid harmonisation is described in Deliverable D2.1. Here, regridding is implemented operationally using variable-type-specific rules so that delivered data remain consistent with the adopted city grid.

Resampling choices depend on the physical meaning of the variable and whether it represents a category, a fraction, or a continuous field. The following default rules are applied:

- Categorical classes (land cover, soil texture): nearest-neighbour resampling for direct reprojection; and majority (mode) aggregation when downscaling from finer to coarser grids.
- Fractional fields (impervious fraction, built-up fraction): aggregation by arithmetic mean when moving from high resolution to the UrbClim grid; values are constrained to physical bounds (e.g., 0-1) after rescaling.

- Continuous geophysical fields (DEM height, anthropogenic heat flux): bilinear resampling for reprojection; optional smoothing filters may be applied to remove isolated artefacts introduced by resampling.
- Vector layers (coastlines, river networks, administrative masks): rasterisation to the target grid using variable-specific rules (binary masks, distance transforms).

For meteorological inputs, the implemented operations consist of city-based ERA5 subsetting, horizontal interpolation, vertical interpolation to UrbClim output heights, and temporal resampling to the model forcing step. Only those resampling choices that affect the delivered files are documented in the relevant dataset subsections of Section 3.

2.5 Quality control and filtering

The minimum implementation-level quality checks retained across D2.2 dataset groups include:

- Spatial integrity: CRS and grid parameters match the city grid index; domain extents and centre points are consistent across all layers.
- Range and physical bounds: variables constrained to physically meaningful limits (0-1 fractions; non-negative precipitation; 0-100% relative humidity).
- Class integrity: land-use and soil-texture classes restricted to the expected set after reclassification; coastal water pixels labelled consistently as sea vs inland water.
- Temporal integrity (meteorological forcing): time series cover the requested simulation period without gaps; units and sign conventions match UrbClim expectations.
- Diagnostics: automatic generation of per-city diagnostic plots and intermediate rasters that allow rapid visual inspection of key variables (domain plots, parameter mosaics, time series and profile checks).
- Logging and provenance: scripts write per-city log files containing command-line arguments, input file versions and processing steps; configuration files are archived with the outputs.

Product-specific quality checks are described in Section 3 where they affect the delivered datasets.

2.6 Uncertainty characterization for tailored input

Source-product uncertainty and dataset validation background are documented in Deliverable D1.3, while the broader project-level validation and uncertainty framework is described in Deliverable D2.1. The purpose of D2.2 is narrower: it records the tailoring-induced uncertainty and implementation-level evidence relevant to the delivered Task 2.2 datasets.

- Inherited uncertainty: the primary uncertainty characteristics of the source product (as reported and summarised in D1.3).
- Processing-induced uncertainty: potential impacts of grid harmonisation and transformations, including resampling artefacts, smoothing and masking decisions.

- Quantitative metrics where available: e.g., inter-model spread for CMIP6 future climate; calibration residuals for GeoDynamiX; summary statistics and plausibility ranges for EO-derived parameters.
- Propagation strategy: how uncertainty information is exposed to, or can be sampled by, downstream model runs.

3 TAILORED DATASETS DELIVERED UNDER DAT1

This section documents the main dataset groups covered by D2.2. For each group, it records the implemented tailoring chain, key inputs, delivered outputs and formats, and the main quality-control and uncertainty information retained for traceability.

3.1 EO-derived terrain inputs for UrbClim

The scientific rationale for using EO-derived surface parameters in UrbClim is described in Deliverable D2.1, while source datasets and their metadata are catalogued in Deliverable D1.3 and section 2.2. This subsection therefore focuses on the Task 2.2 implementation that converts those sources into model-ready UrbClim surface inputs. In CAIAC, EO and ancillary datasets are processed through the terrain workflow implemented in the python script terrain2urbclim.py, which produces the preprocessed land-use raster, the surface NetCDF and the accompanying diagnostics.

3.1.1 Inputs and data harmonisation

All raster inputs are projected to the city-specific UTM coordinate reference system (EPSG) and resampled to the UrbClim grid defined in Section 2.3.
Table 2 summarises the primary EO and ancillary datasets used to derive UrbClim terrain parameters. Most source datasets are documented in the CAIAC Data Inventory: ESA CCI Land Cover (D1.3, Section 2.1), ESA WorldCover (D1.3, Section 2.2), GHSL built-up and population layers (D1.3, Section 2.3), Sentinel-2 NDVI (D1.3, Section 2.5), Copernicus Global DEM (D1.3, Section 2.9), and CMIP6/ERA5 atmospheric datasets (D1.3, Sections 3.2–3.3). Auxiliary datasets used only in the processing chain, such as GSHHS (Section 2.2.3), GISA (Section 2.2.1), AHF (Section 2.2.2) and soil texture maps (Section 2.2.9) are described in this document.

Table 2 EO and ancillary datasets used in the UrbClim terrain-processing chain.

Dataset / layer	Nominal resolution	Role in processing chain	Key operation(s)	Output parameter(s)
GDX land-use map (urban growth chain) derived from ESA CCI Land Cover	300 m	Baseline land-use with scenario consistency	Mode regridding if dxy=600; targeted corrections	Preprocessed land-use
ESA WorldCover land cover	10 m	Refinement for airport pixels and wetland subclassing	Majority aggregation to UrbClim grid; class mapping	Airport replacement class; wetland refinement
Coastline / land-sea mask (GSHHS)	200 m	Separation of inland water vs ocean for coastal domains	Rasterisation; flood-fill classification	Sea-labelled water pixels

Monthly NDVI mosaics	10 m	Vegetation cover fraction (seasonal)	Projection; clipping; scaling to vegetation fraction	COV (12 monthly vegetation fractions)
Digital Elevation Model (DEM)	30 m	Terrain height and slope-related metrics	Reprojection; bilinear resampling, negative elevation clipping	DEM
Soil sealing / impervious fraction	30 m	Impervious surface fraction per grid cell	Aggregation (mean); rescaling; outlier correction	SOIL
GHSL built-up surface + built-up height	10 m	Urban aerodynamic roughness	Aggregation (mean); z0m recomputation	Z0M (urban classes)
Anthropogenic heat flux	500 m	Static urban heat source term	Reprojection; resampling	Anthropogenic Heat Flux (AHF)
Soil texture map	250 m	Hydraulic parameter fields for soil water balance	Nearest-neighbour resampling; class-to-parameter mapping; smoothing	ETAS, PSIS, KSAT, BCH

3.1.2 Land-use preprocessing and class harmonisation

The baseline land-use layer for UrbClim is derived from the GeoDynamiX (GDX) land-use map, which ensures scenario consistency between the urban growth and urban climate chains. Prior to conversion to UrbClim land-use classes, a set of targeted refinements is applied to address known class artefacts and to improve consistency in the urban core and coastal zones. In the operational CAIAC workflow, these preprocessing and parameter-derivation steps are implemented in `terrain2urbclim.py`

3.1.2.1 Airport class replacement using ESA WorldCover

Airport is not a class supported in the UrbClim model; therefore, those pixels are corrected. ESA WorldCover (D1.3, Section 2.2) is aggregated to the UrbClim grid using majority resampling, mapped to the UrbClim land-use taxonomy, and used as an override mask to update airport areas in the preprocessed land-use map.

3.1.2.2 Inland water vs ocean discrimination

For coastal domains, water pixels are separated into inland water and sea to ensure correct treatment in UrbClim. A coastline or land-sea mask (e.g., GSHHS, Section 2.2.3) is rasterised to the city grid and combined with the land-use water class to label coastal water pixels consistently. A flood-fill approach is used to identify connected sea regions where required.

3.1.2.3 Conversion from GDX classes to UrbClim land-use classes

After refinements, the land-use map is converted from GDX classes to UrbClim land-use classes through a deterministic class mapping.

Table 3 provides the mapping used in DAT1. The resulting preprocessed land-use raster is written as a GeoTIFF on the UrbClim grid and serves as the categorical driver for the UrbClim surface parameter lookup tables.

Table 3 Conversion from GDX land-use classes to UrbClim land-use classes.

GDX class	UrbClim class	Refinement step
1 Bare soil	6 Bare soil	-
2 Grassland	7 Grassland	-
3 Cropland	8 Cropland	-
4 Shrubland	9 Shrubland	-
5 Woodland	10 Woodland	-
6 Broadleaf	11 Broadleaf trees	-
7 Needleleaf	12 Needleleaf trees	-
8 Wetland	7 Grassland / 10 Woodland	Herbaceous wetland dominant -> Grassland; Mangrove dominant -> Woodland
9 Urban	1 Urban	-
10 Sub-urban	2 Sub-urban	-
11 Water	14 Inland water / 15 Sea water	Coastline mask: over land -> inland water; over sea -> sea water
12 Airport	1 Urban / 6 Bare soil	Replace airport grid cells by dominant WorldCover-based class

Visualizations of all processed UrbClim land-use are provided in **ANNEX 3: URBCLIM EO PROCESSED DATA**.

3.1.3 Derivation of terrain parameters for UrbClim

Using the preprocessed land-use raster as baseline, the processing chain derives the remaining UrbClim surface parameters. The processing chain has been already described in detail in D2.1, the main derivation steps implemented are summarised below.

3.1.3.1 Vegetation cover fraction from NDVI

Monthly NDVI mosaics are projected to the city UTM grid and converted to vegetation cover fraction. The result is a set of 12 monthly vegetation fractions (COV) per city. Values are constrained to the range 0-1 and masked over water where applicable.

3.1.3.2 Terrain height (DEM)

A Digital Elevation Model is reprojected to the city UTM grid using bilinear resampling and clipped to the modelling domain. Negative elevation artefacts introduced by resampling (e.g., near coastlines) are clipped to zero. The resulting DEM field is used directly by UrbClim and is also used in ancillary processing steps such as lapse-rate adjustment and hydrological index derivation.

3.1.3.3 Soil sealing (impervious surface fraction)

Impervious surface fraction is aggregated to the UrbClim grid by averaging. Source-specific scaling factors are applied to ensure values represent a 0-1 fraction. Extreme outliers (values greater than 1 after rescaling) are corrected using a local neighbourhood statistic. The resulting field (SOIL) represents the fraction of sealed surface per grid cell.

3.1.3.4 Urban roughness length (Z0M)

In built-up areas, the aerodynamic roughness length for momentum is recomputed from building fraction (built-up surface), building height and the vegetation fraction of the peak-greenness month. The combined roughness is clipped to the range 0.3-2 m and replaces the default lookup-table roughness values over urban classes.

3.1.3.5 Anthropogenic heat flux (AHF)

An anthropogenic heat flux raster is reprojected and resampled to the UrbClim grid to provide a constant source term in the UrbClim surface energy balance.

3.1.3.6 Soil texture and hydraulic parameters

A categorical soil texture raster is resampled to the UrbClim grid using nearest-neighbour resampling. Soil hydraulic parameters (porosity ETAS, saturation matric potential PSIS, saturated hydraulic conductivity KSAT and pore-size distribution index BCH) are assigned by texture class and smoothed with a weak Gaussian filter to remove isolated artefacts introduced by resampling.

3.1.4 Output products and file formats

The terrain-processing chain produces a preprocessed land-use GeoTIFF at the UrbClim grid resolution and a NetCDF surface file that is ingested directly by UrbClim. The NetCDF file contains the UrbClim grid and all terrain parameters required by the model, including monthly vegetation fraction.
The NetCDF surface file includes the variables define in the table below:

Table 4 Inventory of the main variables written to the UrbClim surface NetCDF.

Variable	Description	Units / dimensions
X, Y	Grid coordinates in the city projection	m; (ny, nx)
LAT, LON	Geographic coordinates	degrees; (ny, nx)
DEM	Terrain height	m; (ny, nx)
LAN	UrbClim land-use class	1-15; (ny, nx)
LANDMASK / SEAMASK	Binary land / sea mask	0-1; (ny, nx)
COV	Monthly vegetation cover fraction	%; (12, ny, nx)
SOIL	Soil sealing / impervious fraction	%; (ny, nx)
Z0M	Roughness length for momentum	m; (ny, nx)
ALBV, ALBS	Vegetation and ground albedo	0-1; (ny, nx)
EPSV, EPSS	Vegetation and ground emissivity	0-1; (ny, nx)
RST, RP, ROOTB	Vegetation resistance and rooting parameters	s/m, m; (ny, nx)
AHF	Anthropogenic heat flux	W/m²; (ny, nx)
ETAS, PSIS, KSAT, BCH	Soil hydraulic parameters	m3/m3, m, m/s, -; (ny, nx)

3.1.5 Quality control and traceability

The processing scripts generate intermediate GeoTIFF rasters in a dedicated logging directory, enabling inspection of each processing step. As part of the workflow, a diagnostic figure is produced to visualise key terrain parameter fields (vegetation cover, land use, roughness length, terrain height, soil sealing and AHF) for rapid sanity checking.

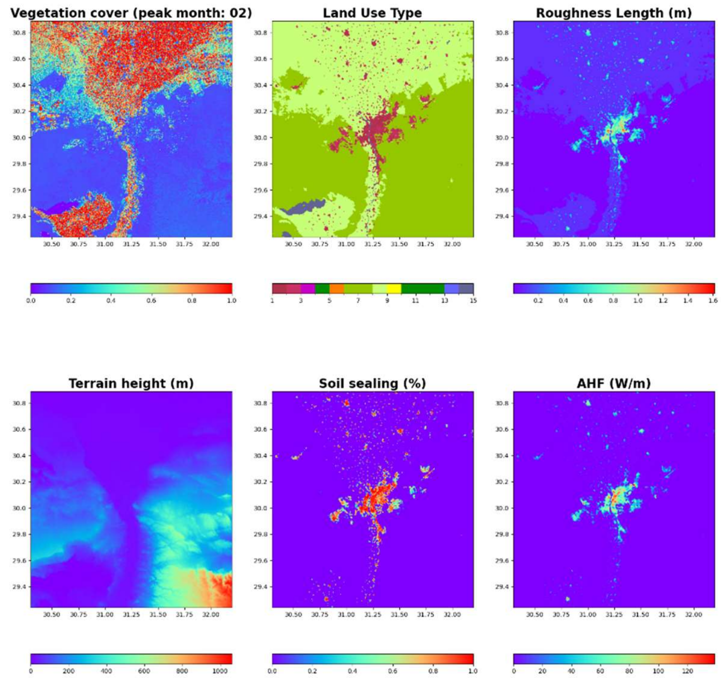


Figure 2 Example of UrbClim terrain file visualisation for Cairo.

3.2 Atmospheric forcing for UrbClim

The role of ERA5 and CMIP6 in the CAIAC urban-climate workflow is described in Deliverable D2.1, while the source-dataset metadata are documented in the CAIAC Data Inventory: ERA5 Atmospheric Reanalysis in D1.3, Section 3.2, and CMIP6 Climate Projections in D1.3, Section 3.3. This subsection documents only the operational Task 2.2 forcing workflow and the outputs it produces. Present-day UrbClim forcing is generated through python scripts cut_ERA5.py and meteo2urbclim_ceda.py, while future-forcing preparation is implemented through make_daily_timeseries.py, delta_matrix_multi_pkl_parser.py and run_all_deltas.py.

3.2.1 Present-day meteorological forcing pipeline (ERA5 to UrbClim)

The present-day forcing workflow consists of two sequential processing steps executed per city and per simulation year. First, cut_ERA5.py extracts compact ERA5 subsets around each city. Second, meteo2urbclim.py converts the extracted fields into a single UrbClim forcing NetCDF file.

3.2.1.1 ERA5 subset extraction

The cut_ERA5.py script extracts a compact subset of ERA5 data around each target city. The extraction domain is defined as a latitude-longitude window centred on the city and large enough to support the subsequent interpolation to the UrbClim grid.
Both ERA5 model-level (3D) fields and ERA5 surface/soil (2D) fields are processed. Files are subset using CDO spatial selection, concatenated into yearly time series, and split into variable-specific NetCDF files.

3.2.1.2 Conversion to UrbClim forcing

The meteo2urbclim_ceda.py script reads the city configuration (centre coordinates, grid dimensions, vertical levels and model time step) and converts the extracted ERA5 subsets into UrbClim-ready forcing files.

For each year, 3D ERA5 fields (u-wind speed, v-wind speed, temperature and specific humidity) and 2D fields (surface pressure, geopotential, precipitation, radiative fluxes, soil temperature and soil moisture) are read and merged. The script removes duplicate latitude/longitude coordinates that can occur for cities located on archive tile boundaries and harmonises longitude conventions where required. Variables are then interpolated to the UrbClim domain centre (clon, clat).

Virtual temperature is computed from temperature and specific humidity. Full-level pressure and geopotential are reconstructed using ECMWF hybrid coefficients (a/b parameters) provided as auxiliary input, and potential temperature is derived from temperature and pressure. The resulting vertical profiles are transformed to geometric height above ground and linearly interpolated to the UrbClim output heights.

Units are harmonised to UrbClim conventions: precipitation in mm, surface pressure in hPa, specific humidity in g/kg, and radiative fluxes in W m-2.

Initial soil temperatures and soil moisture contents are derived from ERA5 soil layers. Soil moisture is recalculated to match the soil texture prescribed for UrbClim using Clapp-Hornberger relations, followed by light spatial smoothing to remove isolated artefacts introduced by resampling. Water-body temperature is set to a proxy value derived from ERA5 soil temperature, while sea-surface temperature is extracted as an area-mean around the city where applicable.

For each city and year, the output is a single NetCDF forcing file containing:

- centre-point atmospheric profiles of u-wind speed, v-wind speed, temperature, potential temperature and specific humidity on UrbClim heights;
- time series of surface pressure, rainfall and downwelling short- and longwave radiation;
- grid-based initial states for soil temperature and soil moisture, together with vegetation and surface water storage variables.

The processing chain produces diagnostics that should be stored alongside the forcing files:

- a domain visualisation (2D and 3D input sampling points) to confirm correct spatial selection.
- time series plots of surface forcing variables to detect temporal gaps or unit issues.
- random vertical profile plots (e.g., potential temperature) to verify vertical interpolation consistency (see figure below).

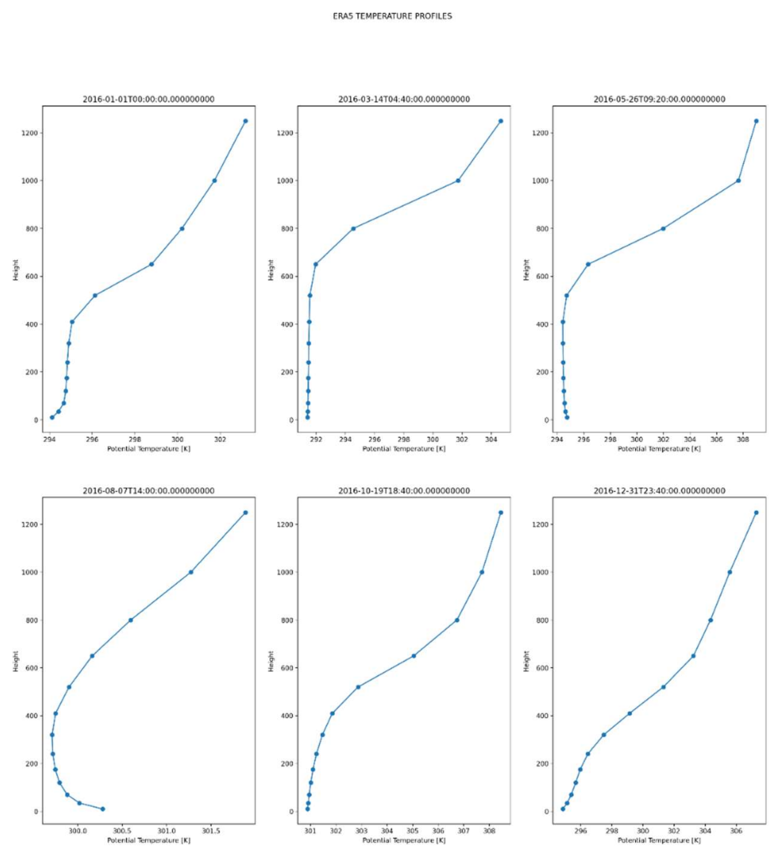


Figure 3 Potential temperature profile randomly sampled over Praia in 2016. Note the unstable profile in the bottom left plot.

3.2.2 Future climate forcing preparation (CMIP6 delta method)

Future meteorological forcing is derived from bias-corrected CMIP6 projections using a percentile-based delta method. In the present CAIAC workflow, the main supporting scripts are make_daily_timeseries.py for baseline indicator extraction and percentile attribution, delta_matrix_multi_pkl_parser.py for percentile-by-month CMIP6 delta matrices, and run_all_deltas.py for batch execution across scenarios and target periods on the HPC infrastructure.

3.2.2.1 Overview

The delta-method implementation can be summarised in three steps:

1. extract a baseline daily indicator time series from present-day UrbClim simulations and assign each day to a monthly percentile class.
2. compute percentile-by-month delta matrices from bias-corrected CMIP6 data for a given scenario and target period.
3. apply the deltas to the baseline indicators to construct a future realisation time series.

3.2.2.2 Baseline indicators and percentile attribution (make_daily_timeseries.py)

The percentile-based delta method requires a mapping between each day in the baseline period and the corresponding percentile class of the baseline distribution, computed separately for each calendar month. In CAIAC, this mapping is derived from UrbClim baseline simulations forced by ERA5 and is evaluated at a representative rural reference location to minimise the influence of urban land-surface feedback.

The script selects a rural reference grid cell based on the spatial 10th percentile of baseline near-surface temperature, after applying a lapse-rate adjustment using the UrbClim DEM and excluding water bodies using the land-water mask. Daily indicators are then computed at this location for the full baseline period: daily minimum and maximum 2 m air temperature, daily mean 2 m air temperature, daily mean wind speed, daily mean land-surface temperature and daily mean relative humidity. Relative humidity is computed from specific humidity, air temperature and an estimated mean surface pressure derived from city mean elevation.

For each calendar month, the baseline distribution of each indicator is summarised using percentiles from 5 to 95 with a 10-percentile step. Each day is assigned to the closest percentile class, yielding a day-by-day percentile index used for delta application. CSV product is produced per city: rural_timeseries_with_percentiles.csv.

3.2.2.3 CMIP6 delta matrices (delta_matrix_multi_pkl_parser.py)

Scenarios are extracted from a bias-corrected CMIP6 archive available on the processing system. For each CMIP6 model and ensemble member, the script computes a delta matrix for a set of source variables relevant for the baseline indicators. The delta matrix is defined on a grid of (percentile, month) using the same percentile set as the baseline attribution and the 12 calendar months.

For each (month, percentile) bin, the script reads pre-computed rolling-window climatology NetCDF files and spatially interpolates the values to the target city location. A baseline climatology is computed over the baseline period, and a future climatology is computed over a five-year window centred on a target year (year-2 to year+2). The delta value is computed as the difference between future and baseline climatology means for the selected bin.

Delta matrices are written to Excel workbooks, with one sheet per variable. Per-model workbooks are produced first, followed by an aggregated multi-model mean and standard deviation across all available models and ensemble members (excluding missing or empty matrices).

3.2.2.4 Multi-model aggregation and uncertainty

To support uncertainty characterisation, the delta workflow retains both the multi-model mean delta matrix and the inter-model spread (standard deviation) for each variable. This information can be propagated to downstream UrbClim simulations through sensitivity testing (e.g., applying mean +/- 1 sigma deltas) or by sampling deltas across models/ensembles.

3.2.2.5 Application of deltas to construct future indicator sequences

Baseline indicators are mapped to CMIP6 source variables, then we apply the following workflow.

Let $X(d)$ denote a baseline daily indicator value for day d and let $p(d)$ be the percentile class assigned to that day for its calendar month $m(d)$. The delta workflow provides a delta matrix $DeltaX(p, m)$ derived from CMIP6 for the chosen scenario and target period. The perturbed (future) indicator sequence is then constructed as:

$$X\_future(d) = X\_baseline(d) + DeltaX(p(d), m(d)).$$

In this formulation, the temporal sequencing of baseline weather is preserved while the distribution is shifted according to CMIP6-derived deltas. Post-processing constraints are applied for bounded or strictly non-negative variables (e.g., relative humidity limited to 0-100%, wind speed constrained to be non-negative).

3.2.2.6 Batch execution across scenarios and target periods (run_all_deltas.py)

The run_all_deltas.py helper script automates delta-matrix computation across multiple scenarios and target periods as defined in the city configuration file. It derives the baseline-period length (20 years) and constructs, for each target period centre year, a corresponding window of equal length centred on that year. Delta matrices are computed at a 5-year step within this window. For each (scenario, target year) combination, the script generates a SLURM job file, activates the project Python environment and runs delta_matrix_multi_pkl_parser.py with the required command-line arguments.

3.2.2.7 Output products and quality assurance

The future-preparation workflow produces intermediate products per city intended to be delivered as transparency artefacts supporting the construction of future forcing and uncertainty quantification:

1. rural_timeseries_with_percentiles.csv (baseline indicators and percentile classes).
2. per-model delta-matrix workbooks ({city}_{year}_{scenario}_{model}_{ensemble}.xlsx).
3. aggregated workbooks containing mean and standard deviation matrices ({city}_{year}_{scenario}_mean.xlsx, with standard deviation sheets suffixed _std).

Quality checks include verifying complete baseline coverage without gaps, checking percentile attribution consistency per month and confirming that delta matrices contain finite values for expected bins.

3.3 GeoDynamiX urban growth inputs/outputs

Urban growth is modelled with GeoDynamiX. A general description of the tool can be found in D1.3, the project configuration, input data and scenario logic in D2.1.

3.3.1 Calibration method

The parameters of the model were calibrated by running the model and analysing the change of land use and population in the 2005-2020 period. This calibration follows usually a semi-automated approach. Optimal parameter values are determined with a Covariance Matrix Adaptation Evolution Strategy (CMA-ES) optimisation algorithm, but the possible ranges of values and the median of the distribution of values in the first generation of the optimisation are preset. Validation on 2020 maps is performed by computing the RMSE of the number of urban land-use cells, the number of suburban land-use cells, and total population in overlapping windows with different resolutions, increasing by a factor of 3. The different resolutions are 900 m, 2.7 km, 8.1 km and 24.3 km (and additionally 72.9 km for large grids). After running the optimisation, there are several tests to inspect if the optimal set of parameter values is final. Parameter values should not be close to minima or maxima. If the model is run in simulation mode towards the future, Zipf's law on the frequency of urban cluster sizes should be respected. A visual inspection is also performed.

The number of cities in the CAIAC project is larger than in earlier applications of the model. Therefore, we selected several pilot cities and further automated the procedure for all other cities. After an initial analysis, 6 cities with clearly distinct growth patterns were chosen as the pilot cities for the calibration algorithm: Bamako, Cairo, Kigali, Niamey, Nouakchott and Port Louis. The initial parameter ranges were copied from earlier projects and iteratively adapted if necessary, after performing the different mentioned tests. For the 48 other cities 4 clean starts (with different random behaviour) of the optimisation runs were performed and compared. If the best calibration result failed the test, then a fallback procedure was started where initial values for the calibration are copied from the final results of the 6 pilot cities. The evaluation then ranks the calibration results of these 6 runs and selects the highest optimisation result that passes the test. This procedure was mainly necessary for some of the smaller cities with limited change during the calibration period.

3.3.2 Simulation inputs and results

The model starts from land-use and population maps for which the interaction rules are calibrated, with suitability maps, zoning maps and accessibility maps as additional criteria. Specific input data for land use, suitability, zoning and accessibility were mentioned in D1.3 and D2.1.

The model is run for each city from 2020 to 2100 for the SSP2 and SSP3 scenarios. Model parameters on land-use interactions are copied from the historical calibration. Urban and suburban land size are computed from global population and urbanisation scenarios per country (see D2.1). All annual intermediate results of land use and population can be saved as GeoTIFF files. The growth of urban land use between 2020 and 2100 in SSP2 for the 6 pilot cities is visualised in Figure 4 and Figure 5.

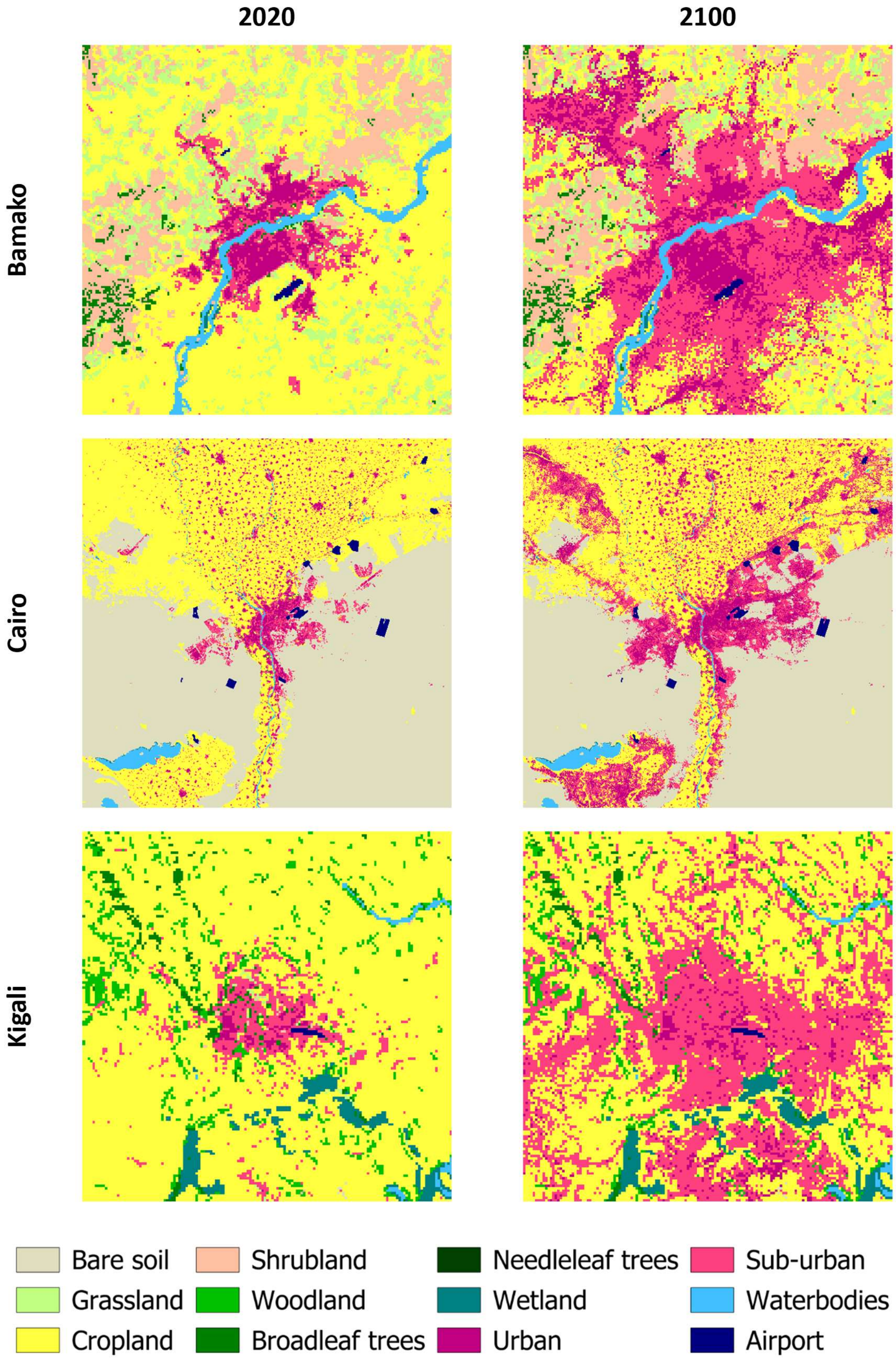


Figure 4. Initial 2020 land-use map and SSP2 simulation result in Bamako, Cairo and Kigali.

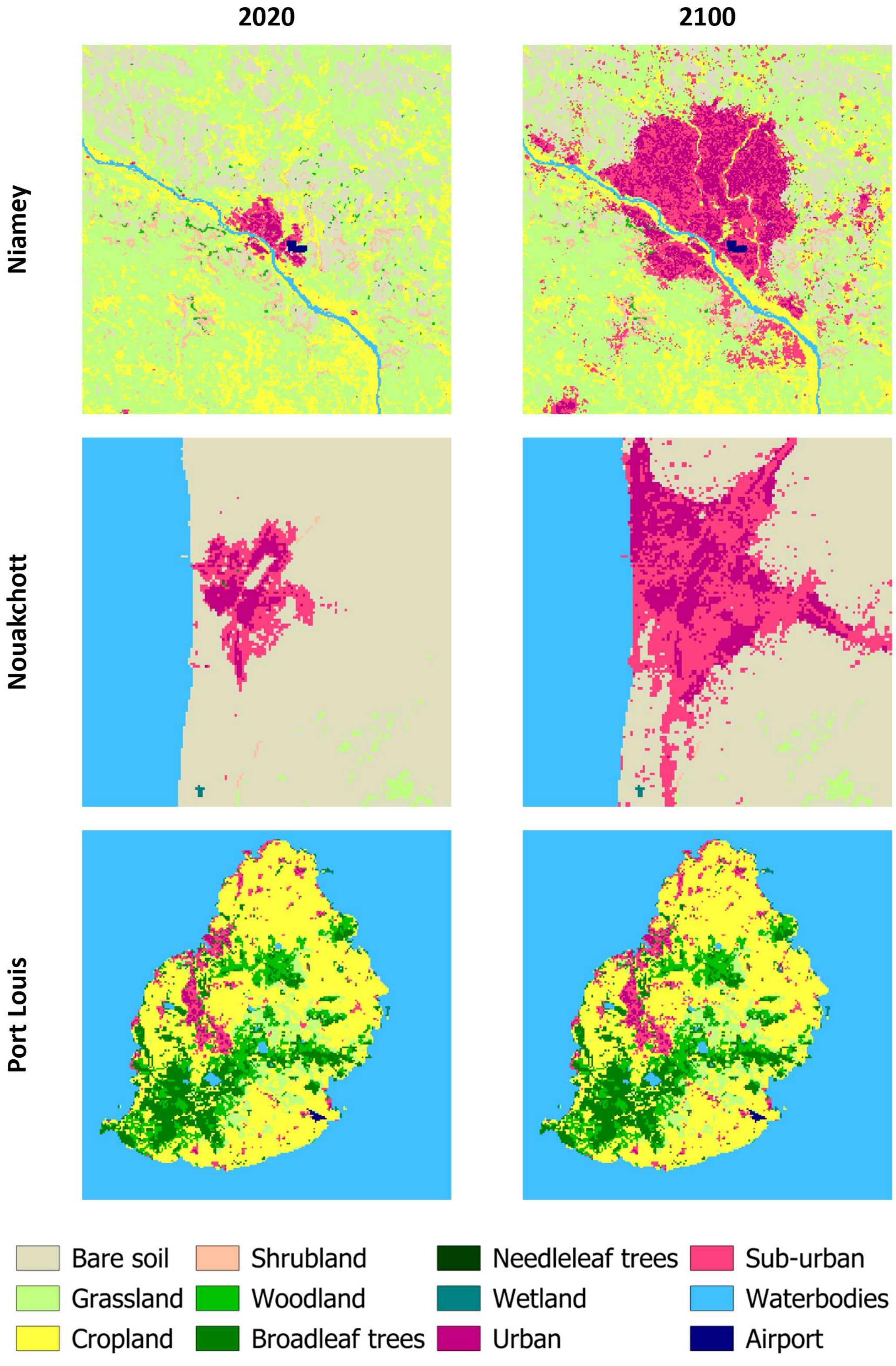


Figure 5. Initial 2020 land-use map and SSP2 simulation result in Niamey, Nouakchott and Port Louis.

3.3.3 Quality control and uncertainty

The quality control of the calibration procedure consists of the mentioned semi-automated procedure including a visual inspection. The uncertainty generated by the input data and demographic scenarios is not directly considered. The spatial uncertainty of the simulation itself is quantified in a Monte Carlo analysis. The potential for change of the model is multiplied by a stochastic factor. This factor is randomly drawn from a highly skewed distribution in which small perturbations are likely, but larger perturbations are possible in few locations. This stochasticity represents the different options to extend the urban area in different directions or build new neighbourhoods. The Monte Carlo analysis (running the stochastic model 100 times) results as such in probability maps for all land uses. A probability map for combined urban land (urban + suburban) was calculated.

The spatial uncertainty of urbanization in the SSP2 scenario of 2100 is shown for the 6 pilot cities in Figure 6.

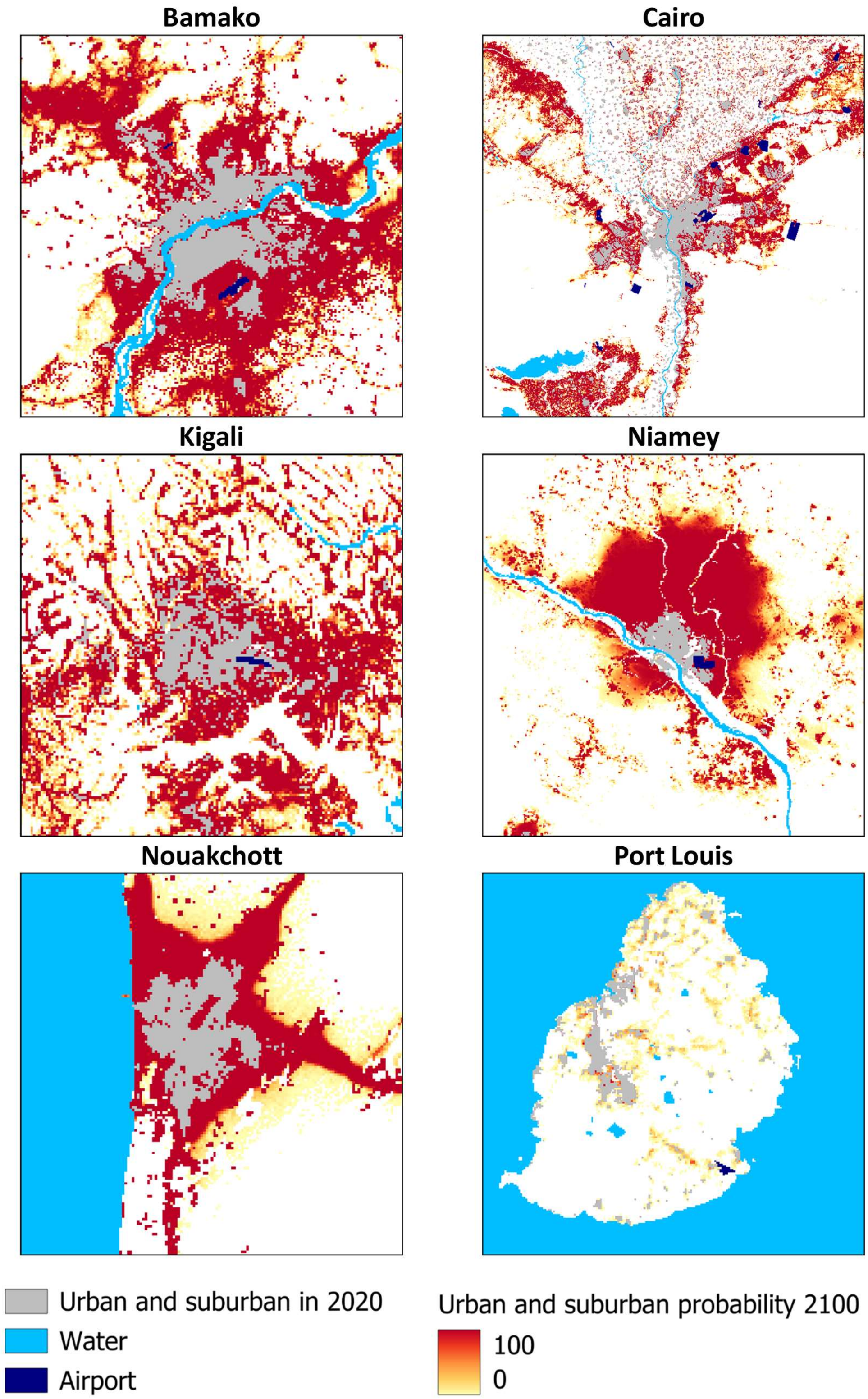


Figure 6. Monte Carlo analysis in the 6 pilot cities, showing spatial uncertainty of urbanisation in SSP2.

3.4 Flood-model inputs (extremes/return periods and supporting layers)

The flood-modelling workflow uses both datasets already documented in the CAIAC Data Inventory and additional flood-specific datasets that are not fully described as standalone entries in D1.3. D1.3 already documents Sentinel-1 Global Flood Monitoring (D1.3 Section 2.4), ESA CCI Soil Moisture (D1.3 Section 2.7), ESA CCI River Discharge (D1.3 Section 2.8), ERA5 (D1.3 Section 3.2), and CMIP6 Climate Projections (D1.3 Section 3.3). Additional datasets used in the flood workflow include TAMSAT precipitation (Section 2.2.7), GRDC river discharge records (Section 2.2.4), ISIMIP3b hydrological simulations (Section 2.2.5), SRTM DEM (Section 2.2.8) and HWSD/SoilGrids soil properties (Section 2.2.9).

3.4.1 Hydrological extremes and return-period products

Pluvial and fluvial urban flooding are primarily associated with extreme surface runoff and extreme river discharge, respectively. For pluvial flooding, the physically based hydrological model SWAT+ is used to generate surface-runoff series from historical and future precipitation and temperature inputs. Future hydroclimatic conditions are represented using bias-adjusted climate scenarios SSP1-2.6, SSP2-4.5, SSP3-7.0 and SSP5-8.5. Prior to hydrological modelling, the climate-model forcing is adjusted against observational and reanalysis reference datasets to reduce systematic biases. The resulting surface-runoff series are subsequently subjected to frequency analysis to estimate quantiles and return periods for extreme pluvial events. For fluvial flooding, river-discharge series are obtained from ISIMIP 3b hydrological simulations and subsequently adjusted to improve consistency with available observed discharge records from the Global Runoff Data Centre (GRDC). Frequency analysis is then applied to the discharge time series to estimate the quantiles and return period of high-flow events.

Return-period surface-runoff and river-discharge values for the 2, 5, 10, 25, 50, 100 and 200 year events are then used in a hydrodynamic modelling workflow based on HEC-RAS to estimate the inundation depth associated with these extreme events. These inundation-depth products form a key component of the Rapid Inundation Labelling (RIL) framework, in which dynamic hydrometeorological variables are combined with static flood-risk characteristics to generate instantaneous inundation maps.

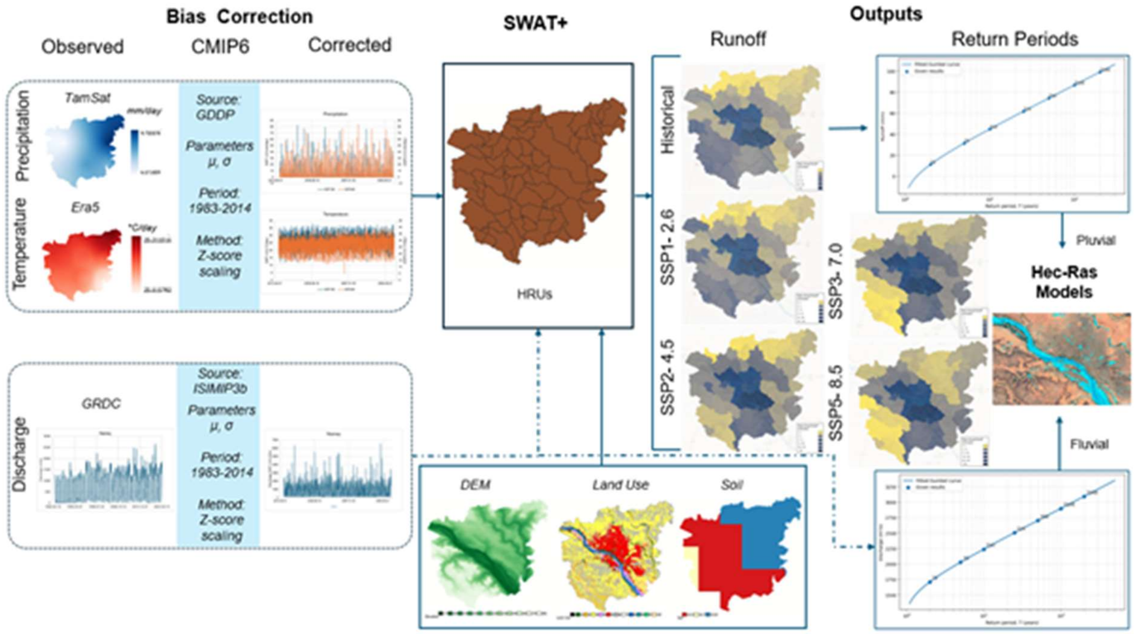


Figure 7 Hydrological modelling framework and simulation result in Niamey.

3.4.1.1 Bias Correction and Z-Score Scaling

Climate model outputs from CMIP6 (GDDP) for precipitation, temperature; and from ISIMIP3b for river discharge, exhibit systematic biases relative to observational records. To remove systematic biases prior to hydrological modelling, a parametric scaling approach based on Z-score standardisation was applied. For a given variable X (precipitation, temperature, or discharge), the bias-corrected value is obtained as:

$$Z_{mod} = \frac{(X_{mod} - \mu_{gcm})}{\sigma_{gcm}}$$
$$\hat{X}_{mod} = \mu_{obs} + \sigma_{obs} \cdot Z_{mod}$$

where μ and σ denote the mean and standard deviation estimated over the calibration period 1983–2014, and the subscripts *obs* and *gcm* refer to the observed reference and modelled series, respectively. This procedure preserves the temporal variability of the model while anchoring its distributional properties to those of the observations.

For precipitation, the observational reference was TAMSAT (0.0375°, 1983–2024; University of Reading, see Section 2.2.7) and ERA5 reanalysis temperature (0.25°, 1981–2024; ECMWF, see D1.3, Section 2.9). For river discharge, the reference was station-based records from the Global Runoff Data Centre (GRDC, see 2.2.4), operated by the German Federal Institute of Hydrology (BfG). The calibration period for all bias correction procedures was set to 1983–2014, ensuring temporal overlap between observational and reanalysis datasets.

An open-source processing pipeline was developed to automate the bias correction workflow, including data ingestion, parameter estimation, transformation, and export of corrected time series in NetCDF format. The codebase is publicly available and documented to facilitate reproducibility across different study regions and climate scenarios.

3.4.1.2. Simulation of Runoff with SWAT+

Surface runoff was simulated using the Soil and Water Assessment Tool (SWAT+), a physically-based, semi-distributed hydrological model. The basin was discretised into Hydrologic Response Units (HRUs), which represent unique combinations of land use, soil type, and slope class, allowing spatially distributed simulation of the water balance. The simulated outputs for each HRU are stored as structured text files (.txt), which contain the daily runoff series for each landscape unit. These files are subsequently combined with the corresponding spatial delineation of each HRU in shapefile format (.shp) to generate gridded NetCDF files at a daily timestep and a spatial resolution of 100 × 100 m, enabling spatially explicit representation of runoff across the basin. River discharge corrected is incorporated as a complementary input to characterise streamflow condition. Below is an example of the pluvial flood map product for an urban area in the city of Niamey, based on runoff simulated by SWAT+ under different climate scenarios.

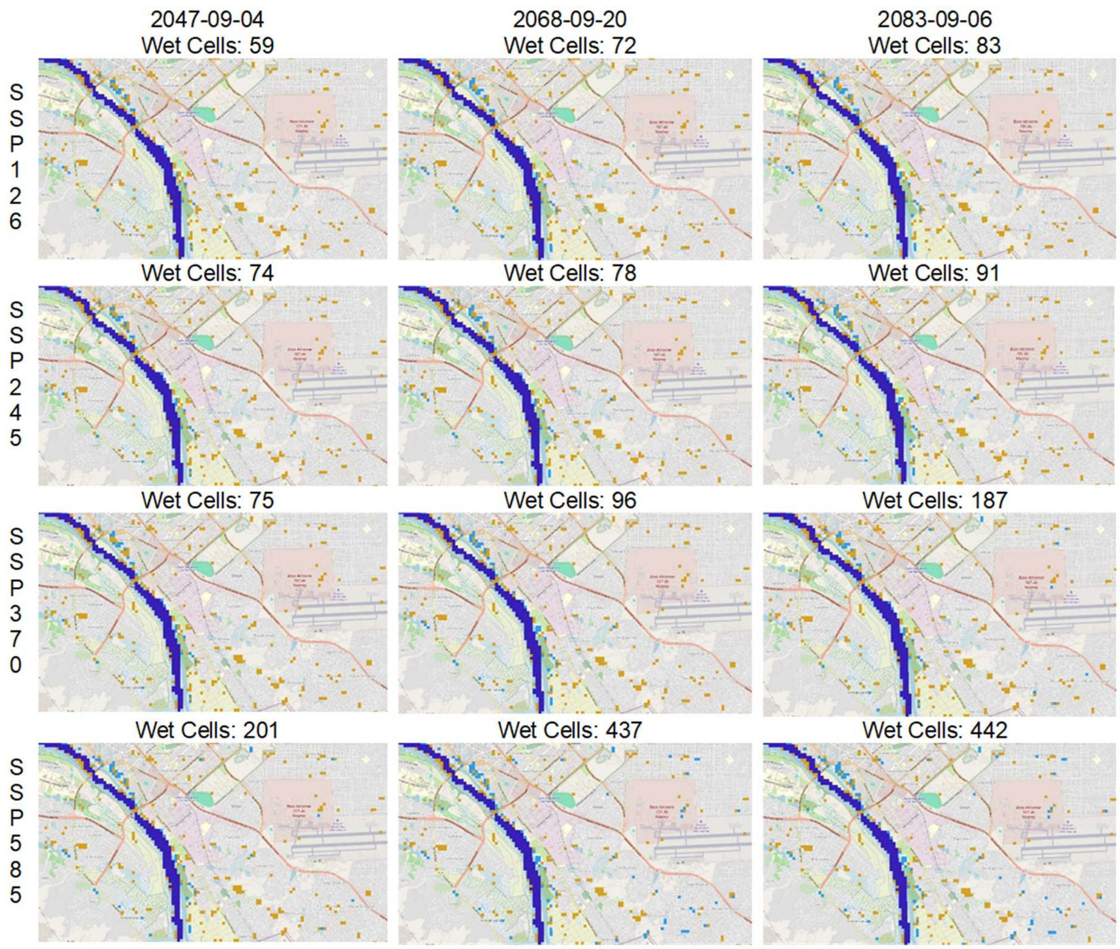


Figure 8 Hydrological modelling framework and simulation result in Niamey.

3.4.1.3 Return Period Estimation via Frequency Analysis

Return periods for runoff and discharge were derived through statistical Frequency Analysis. For each discharge time series and each HRU runoff field stored in NetCDF format, Annual Maximum Series (AMS) were extracted from the corrected daily records by selecting the maximum value per hydrological year. Two complementary approaches were applied: an empirical estimation based on sample quantiles, and a parametric fit using the Gumbel (Extreme Value Type I) distribution.

The Gumbel distribution parameters were estimated by the Method of Moments (MOM):

$$\beta = \sqrt{6} \cdot \frac{\sigma}{\pi}$$

$$\alpha = \mu - \gamma \cdot \beta$$

where μ and σ are the sample mean and standard deviation of the AMS, and $\gamma \approx 0.5772$ is the Euler–Mascheroni constant. Return-period quantiles were then obtained by inverting the fitted distribution:

$$x_T = \alpha - \beta \cdot \ln\left(-\ln\left(1 - \frac{1}{T}\right)\right)$$

for standard return periods T = 2, 5, 10, 25, 50, 100, and 200 years. Goodness-of-fit was assessed through a QQ-based summary including the correlation coefficient between empirical and fitted quantiles (*qq_corr*), RMSE, and bias.

3.4.2 Static supporting layers and ancillary predictors

In addition to climate datasets, static geographic layers such as topography (Section 2.2.8), land cover (D1.3, Section 2.2), and soil properties (Section 2.2.9) were incorporated to support hydrological and hydrodynamic modelling. These datasets were used to define watershed characteristics within the SWAT+ framework, enabling the delineation of watersheds, landscape units (LSUs) and hydrologic response units (HRUs). The SWAT+ model then used these watershed attributes, along with climate inputs, to simulate surface runoff and river discharge, providing key inputs for the flood hazard analysis and the derivation of return period indicators. As for the HEC-RAS, DEM and land-use layers are essential supporting inputs. The DEM provides the terrain representation needed to define flow paths, floodplain storage and inundation depth, while land-use information supports the assignment of spatially distributed surface-roughness parameters (Manning's n), which influence flow velocity, flood extent and water-depth distribution.

Table 5 Inventory of the variables incorporated in the flood products.

Dataset	Variable	Source	Res.	Period	Inventory reference
ERA5	Temperature	ECMWF	0.25°	1981-2024	D1.3, Section 3.2
TAMSAT	Precipitation	Univ. Reading	0.0375°	1983-2024	D2.2, Section 2.2.7
GRDC	Discharge	BafG	Stations	Variable	D2.2, Section 2.2.4

CMIP6/GDDP	Temp+Precip	GDDP	~100 km	2015-2100	D2.2, Section 2.2.6
CIMP6/ISIMIP3b	Discharge	ISIMIP	~50 km	2015-2100	D2.2, Section 2.2.5
DEM	Elevation	SRTM	30m	Static	D2.2, Section 2.2.8
Land Use	Land Cover	ESA World Cover	10m	2020	D1.3, Section 2.2
Soil	Soil Props.	HWSD/SoilGrids	250m	Static	D2.2, Section 2.2.9

3.4.3 Quality control and uncertainty

Quality control for the flood-input package included the inspection of the derived precipitation and temperature time series to identify missing data, verify temporal consistency, and assess whether values remained within physically plausible ranges. In addition, the derived series were compared with alternative reference datasets to check for consistency and to identify potential biases or anomalous behaviour prior to their use in the hydrological workflow. For the inundation outputs, the simulated flood extent and depth patterns were compared with satellite imagery from flood events associated with similar discharge quantiles, in order to evaluate the realism of the modelled inundation and, where necessary, adjust the results so that the simulated and observed flood patterns correspond more closely.

To illustrate the quality and interpretability of the products, Figure 9 below presents several sample rapid inundation labelling (RIL) threshold curves used in the runoff-based flood mapping procedure. Through these relationships, the runoff threshold required to initiate inundation is identified for each cell. For illustration, four representative cells were selected from four different parts of the city of Niamey. It can be observed that, when cells are located close to the river or drainage pathways, as in cell a, lower runoff thresholds are obtained and inundation is triggered more easily. As greater distance from the drainage network is introduced, and as elevation and land-use conditions change, cells with much higher runoff thresholds are identified. This behaviour is illustrated by point c, where inundation is initiated only when runoff reaches approximately 1.066 m. In contrast, a point such as d is characterized by an extremely high threshold of about 39 m, indicating that such cells remain effectively dry and are not expected to become inundated during flood events.

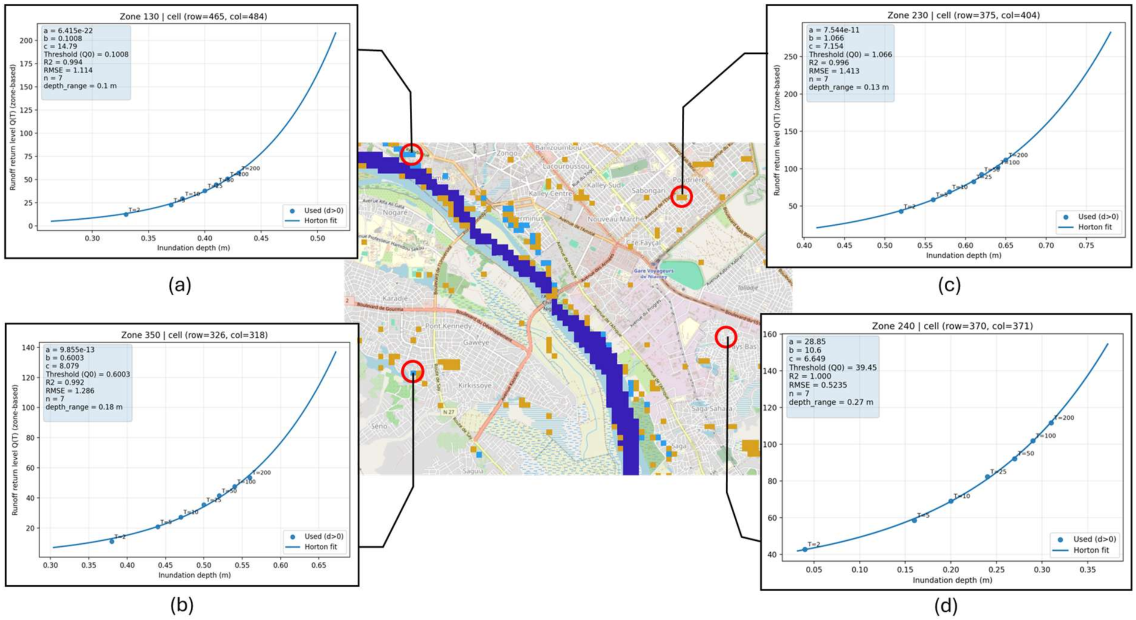


Figure 9 Runoff-based flood mapping procedure applied during an extreme event simulation for Niamey.

To further assess the uncertainty of the threshold maps, the goodness of fit of the RIL framework was evaluated across all fitted cells using the coefficient of determination (R^2). This step was carried out to verify that the majority of fitted relationships were of acceptable quality. As illustrated in Figure 10, depicting the R^2 histogram for climate change scenario SSP5-8.5, 91.42% of the cells showed an R^2 value greater than 0.9, indicating that most fitted models achieved a high level of agreement and providing confidence in the reliability of the generated threshold maps.

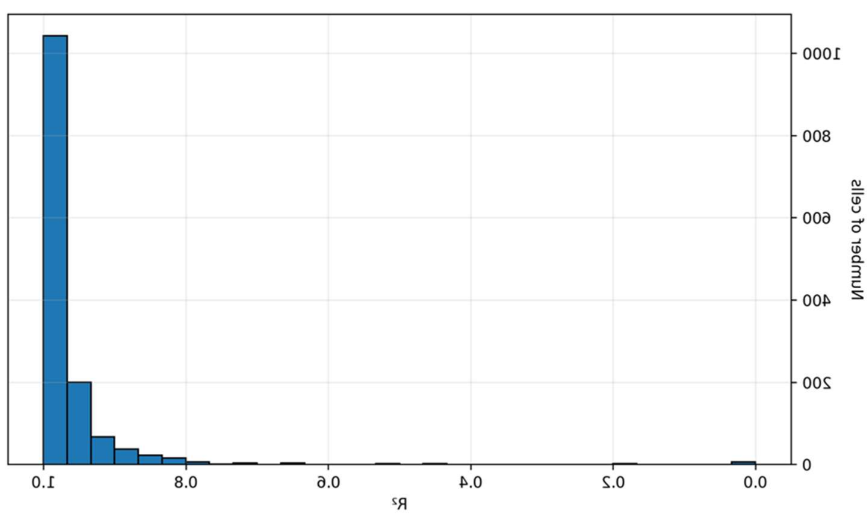


Figure 10 Distribution of RIL method's R^2 across all cells in the case of Niamey's threshold map

4 DELIVERY, ACCESS AND USAGE GUIDANCE

This section clarifies how the tailored datasets documented in D2.2 are made available within the consortium at the present stage of the project. D2.2 itself is a documentation deliverable: it records the processing chains, formats, provenance and access arrangements of the tailored data products generated under Task 2.2. It should therefore not be interpreted as a single central data package through which all consortium partners must exchange data.

In operational terms, the data-sharing arrangement presently follows the needs of the modelling chains. The urban climate workflow (UrbClim) and the urban growth workflow (GeoDynamiX) are implemented on the same VITO high-performance computing infrastructure. As a result, the tailored inputs and outputs needed by these two modelling chains are already accessible within the same processing environment, and no additional inter-partner transfer step is required for routine execution.

The flood-modelling workflow is handled on a separate infrastructure. At the current stage of the project, this workflow does not require direct ingestion of the tailored urban-climate or urban-growth data products in order to run its own processing chain. Consequently, no systematic transfer of the full UrbClim/GeoDynamiX data holdings to the flood-modelling environment is required at present. Where targeted exchange of specific supporting layers or summary information becomes necessary, this can be handled on a case-by-case basis through the agreed consortium coordination channels.

In that sense, consortium-level data sharing under Task 2.2 is already achieved through the existing project infrastructure and partner workflows:

- shared operational access between the urban growth and urban climate teams on the common HPC environment,
- separate but documented handling of the flood workflow on the uOttawa side,
- cross-consortium visibility through this D2.2 note, the associated manifests, configuration files, logs and regular technical exchanges.

Broader publication and open-access dissemination of selected data products and code are foreseen later under WP4 rather than through D2.2 itself.

4.1 Example access snippets

Where products are accessed directly on the VITO processing infrastructure, NetCDF and GeoTIFF files can be read using standard Python geospatial libraries. The following examples illustrate typical access patterns for the UrbClim/GeoDynamiX environment (paths are indicative):

```
import xarray as xr
import rioxarray as rxr

# UrbClim surface NetCDF
ds = xr.open_dataset('urbclim_surface/Kigali/surface_Kigali_300m_v1.nc')
dem = ds['DEM']

# UrbClim forcing NetCDF
fc = xr.open_dataset('urbclim_forcing_present/Kigali/2015/forcing_Kigali_2015_v1.nc')
tas = fc['T']

# Preprocessed land-use GeoTIFF
lu = rxr.open_rasterio('urbclim_surface/Kigali/landuse_preprocessed_Kigali_300m_v1.tif'
```

5 CONCLUSION

This deliverable, D2.2 (DAT1), does not re-specify the full dataset inventory or the scientific methodology of CAIAC; those functions are fulfilled by D1.3 and D2.1 respectively. However, where Task 2.2 uses additional workflow-specific datasets not fully documented in D1.3, this deliverable provides concise inventory-style descriptions to ensure traceability of access points, licensing and uncertainty considerations.

D2.2 records the operational data-tailoring implementation of Task 2.2, including the adopted city grids, the scripts and configuration used to generate UrbClim surface and forcing inputs, the GeoDynamiX exchange datasets relevant to the shared VITO infrastructure, the flood-related data interfaces and support products, and the associated manifests, logs and diagnostics.

Its primary function is therefore traceability and handover of tailored products, not repetition of background methodological material. The document is intended to support consortium use, ESA review, and later publication or extension activities by making clear which artefacts exist, how they were generated, where they are operationally available and which uncertainty or quality-control elements accompany them.

6 REFERENCES

Huang, X., Yang, J., Wang, W., & Liu, Z. (2022). Mapping 10-m global impervious surface area (GISA-10m) using multi-source geospatial data. Earth System Science Data, 14, 3649–3672. doi:10.5194/essd-14-3649-2022.

Wang, S., Hu, D., Yu, C., Wang, Y., & Chen, S. (2022). Global mapping of surface 500-m anthropogenic heat flux supported by multi-source data. Urban Climate, 43, 101175. doi:10.1016/j.uclim.2022.101175.

ANNEX 1: DELIVERED DATASET GROUPS AND MANIFEST TEMPLATE

Table 6 Summary of DAT1 dataset groups and delivered artefacts.

Dataset group	Delivered per city	Format(s)	Main content	Evidence / provenance
Grids and domains	Yes	Excel, KML/GeoJSON	Grid parameters, CRS, bounds, domain polygons	city_grids.xlsx; grid generation script + logs
UrbClim surface	Yes	NetCDF, GeoTIFF	Terrain height, land use, vegetation fraction, roughness, soil, AHF	terrain2urbclim.py config + logs; diagnostic plots
UrbClim forcing (present)	Yes (yearly)	NetCDF	Atmospheric profiles and surface forcing derived from ERA5	cut_ERA5.py + meteo2urbclim.py logs; QA plots
Future forcing preparation	Yes	CSV, Excel	Baseline indicators, percentile attribution, CMIP6 delta matrices (mean and spread)	make_daily_timeseries.py; delta_matrix scripts; SLURM logs
GeoDynamiX	Yes	GeoTIFF	Annual land use and population maps (SSP2/SSP3)	Model workflow scripts and setting files
Flood inputs	City/catchment dependent	NetCDF, GeoTIFF, method note	Return-period precipitation, static predictors, EO flood maps (where available)	Method note; QA checks; uncertainty summaries
Validation	City dependent	NetCDF/GeoTIFF/CSV	LST, SAR flood maps, river discharge/soil moisture (where available)	Product metadata; quality flags; uncertainty notes

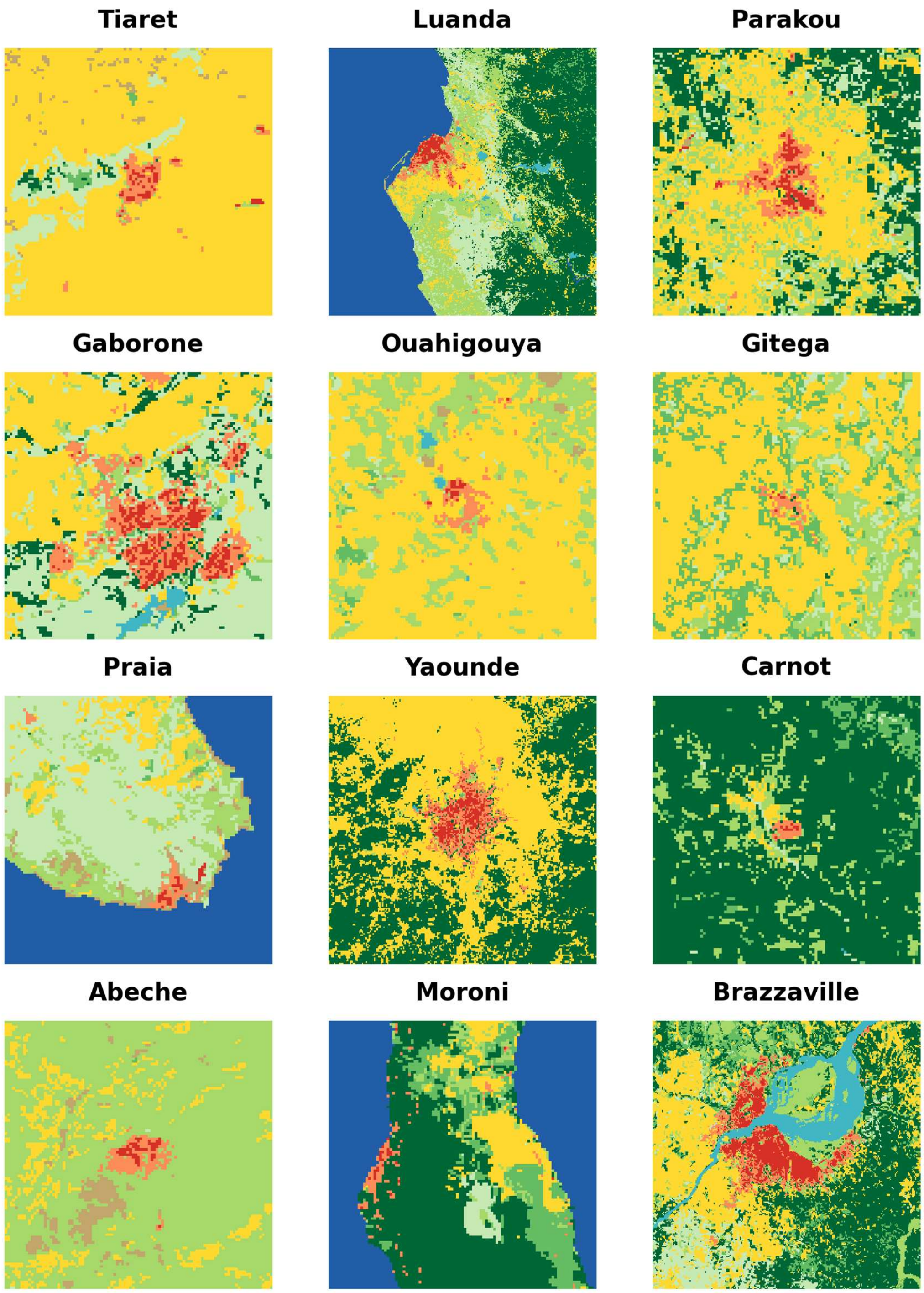
ANNEX 2: CITY GRID

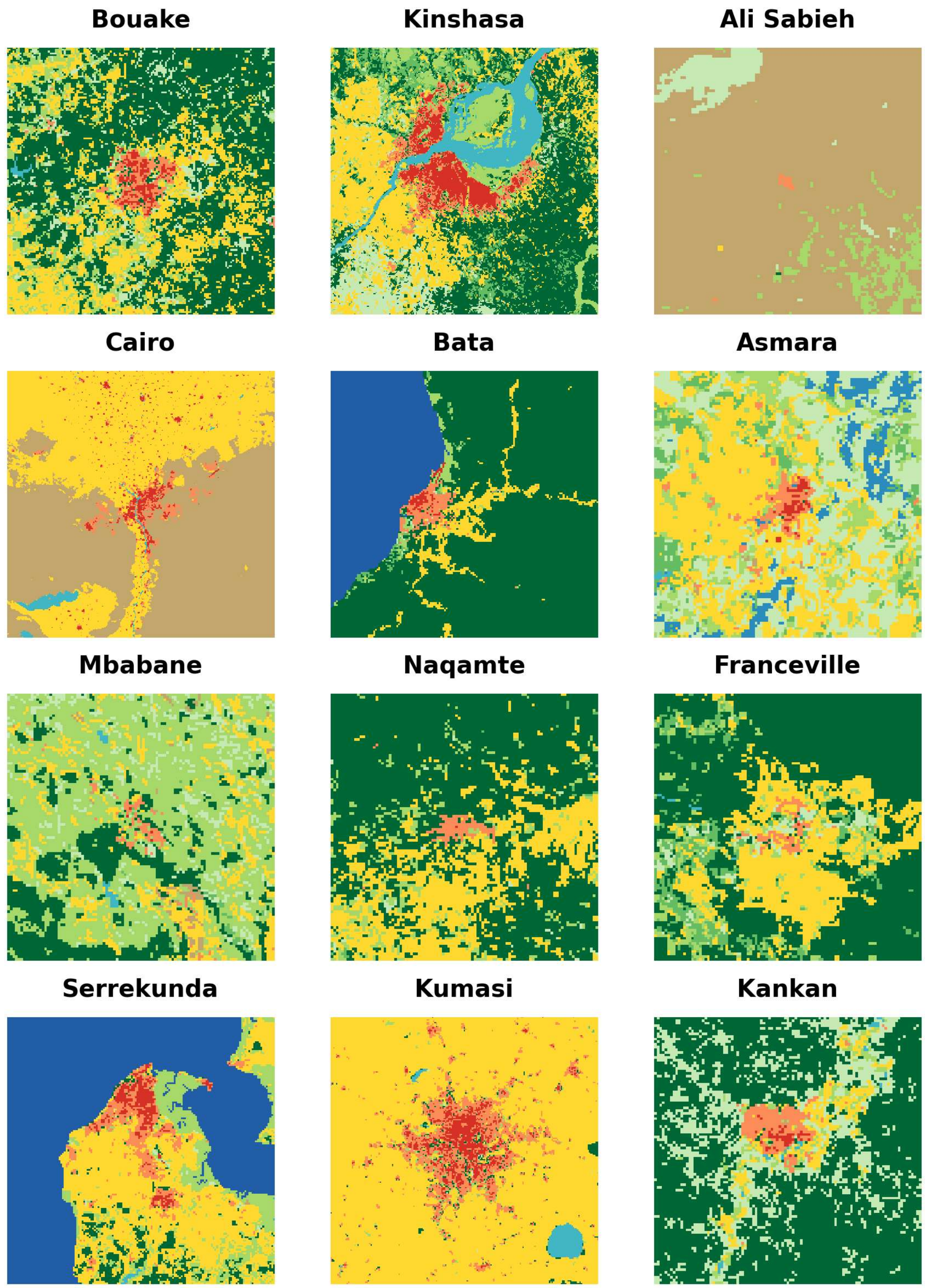
Table 7 City grid parameters (EPSG, resolution and domain size).

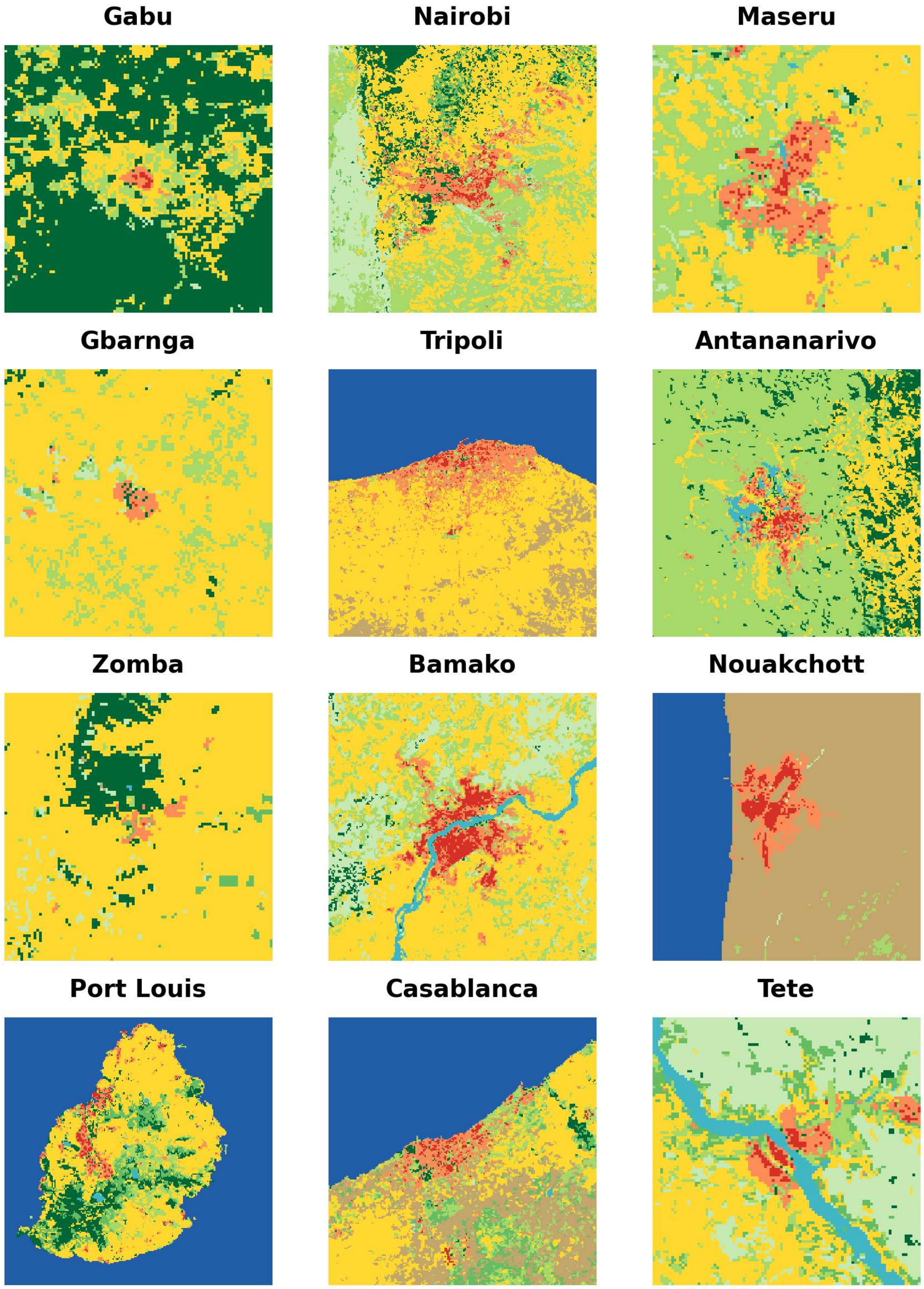
city	epsg	dxy	n	width_km	clon	clat
Cairo	32636	600	301	180.6	31.2542	30.0681
Kinshasa	32733	300	251	75.3	15.3542	-4.3736
Luanda	32733	600	301	180.6	13.4764	-9.0514
Johannesburg	32735	600	301	180.6	28.0403	-26.1819
Khartoum	32636	300	251	75.3	32.5069	15.6069
Nairobi	32737	300	251	75.3	36.8514	-1.2653
Yaounde	32632	300	201	60.3	11.5181	3.8403
Kumasi	32630	300	201	60.3	-1.6153	6.6903
Casablanca	32629	300	201	60.3	-7.5431	33.5819
Antananarivo	32738	300	201	60.3	47.5347	-18.8681
Bamako	32629	300	201	60.3	-7.9792	12.6236
Lusaka	32735	300	151	45.3	28.2903	-15.3986
Brazzaville	32733	300	251	75.3	15.3542	-4.3736
Tunis	32632	300	201	60.3	10.0986	36.7681
Harare	32736	300	201	60.3	31.0431	-17.8819
Nouakchott	32628	300	151	45.3	-15.9375	18.0403
Kigali	32736	300	151	45.3	30.0958	-1.9514
Niamey	32631	300	301	90.3	2.1181	13.5264
Abuja	32632	300	251	75.3	7.4347	9.0681
Serrekunda	32628	300	151	45.3	-16.6819	13.3514
Tripoli	32633	300	251	75.3	13.1819	32.7486
Bata	32632	300	151	45.3	9.8431	1.8542
Bouake	32630	300	151	45.3	-5.0319	7.7014
Port Louis	32740	300	221	66.3	57.5986	-20.2708
Tete	32736	300	101	30.3	33.5931	-16.1569
Windhoek	32733	300	101	30.3	17.0458	-22.5292
Abeche	32634	300	101	30.3	20.8292	13.8264
Parakou	32631	300	111	33.3	2.6208	9.3486
Mbale City	32636	300	101	30.3	34.1931	1.0819
Bo	32629	300	101	30.3	-11.7347	7.9542
Franceville	32733	300	101	30.3	13.5931	-1.6236
Gaborone	32735	300	101	30.3	25.8764	-24.6125
Maseru	32735	300	101	30.3	27.5153	-29.3264
Juba	32636	300	201	60.3	31.5681	4.8542
Kankan	32629	300	101	30.3	-9.2958	10.3653
Dodoma	32736	300	101	30.3	35.7431	-6.1625
Asmara	32637	300	101	30.3	38.9292	15.3319
Tiaret	32631	300	101	30.3	1.3236	35.3625
Naqamte	32637	300	101	30.3	36.5569	9.0903
Galkayo	32638	300	101	30.3	47.4264	6.7708
Sokode	32631	300	101	30.3	1.1431	8.9875
Gitega	32735	300	101	30.3	29.9236	-3.4292
Moroni	32738	300	101	30.3	43.3403	-11.6931
Praia	32627	300	101	30.3	-23.5514	14.9847
Sao Tome	32632	300	101	30.3	6.6597	0.2903
Ouahigouya	32630	300	101	30.3	-2.4153	13.5625
Mbabane	32736	300	101	30.3	31.1319	-26.3181
Carnot	32633	300	101	30.3	15.8764	4.9431
Ali Sabieh	32638	300	101	30.3	42.7069	11.1542
Gabu	32628	300	101	30.3	-14.2319	12.2847
Victoria	32740	300	101	30.3	55.4736	-4.6792
Gbarnga	32629	300	101	30.3	-9.4681	6.9986
Zomba	32736	300	101	30.3	35.3208	-15.4014
Darou Mousty	32628	300	101	30.3	-16.0486	15.0403

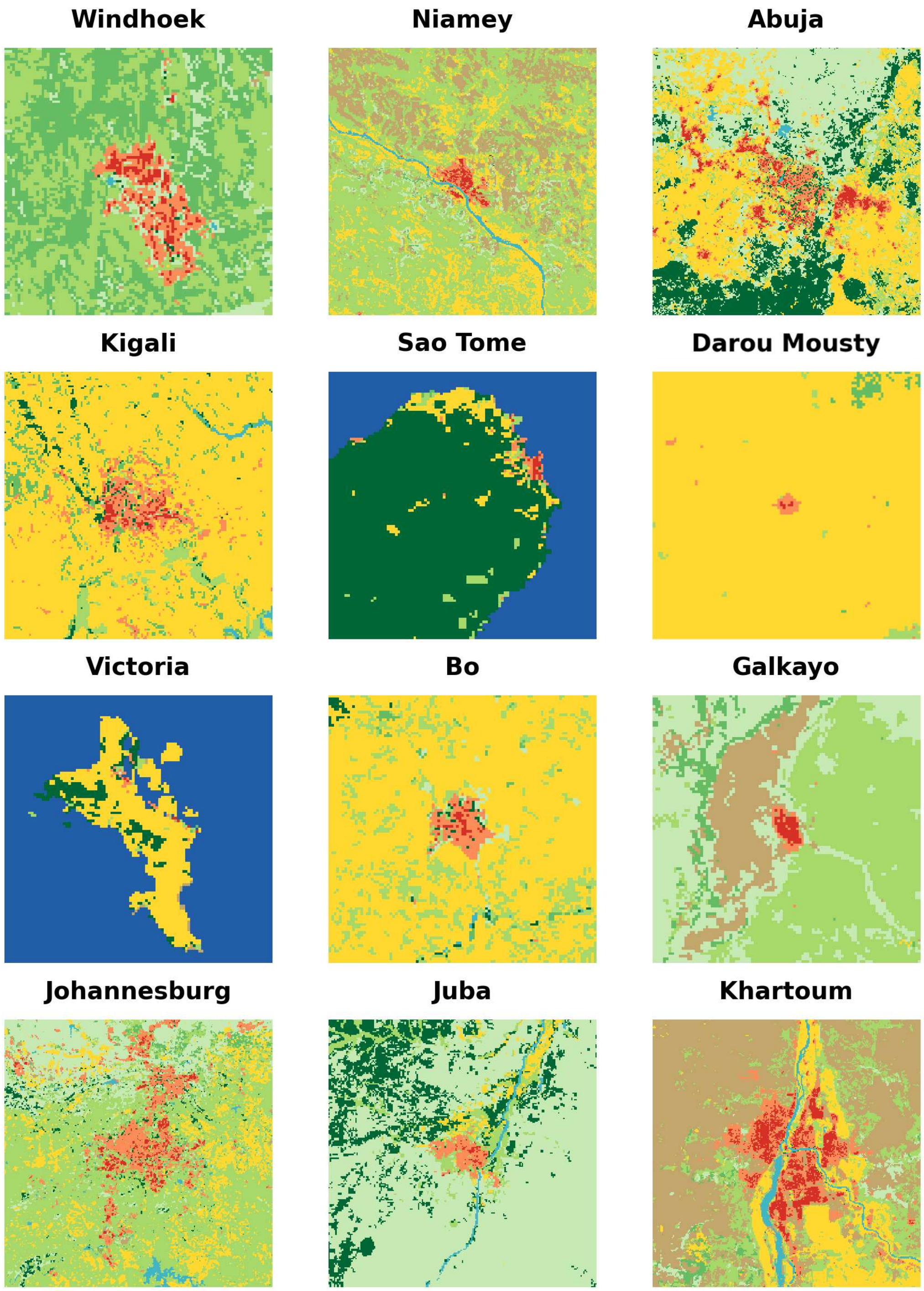
ANNEX 3: URBCLIM EO PROCESSED DATA

Land cover




Bouake
Kinshasa
Ali Sabieh
Cairo
Bata
Asmara
Mbabane
Naqamte
Franceville
Serrekunda
Kumasi
Kankan


Gabu
Nairobi
Maseru
Gbarnga
Tripoli
Antananarivo
Zomba
Bamako
Nouakchott
Port Louis
Casablanca
Tete


Windhoek
Niamey
Abuja
Kigali
Sao Tome
Darou Mousty
Victoria
Bo
Galkayo
Johannesburg
Juba
Khartoum

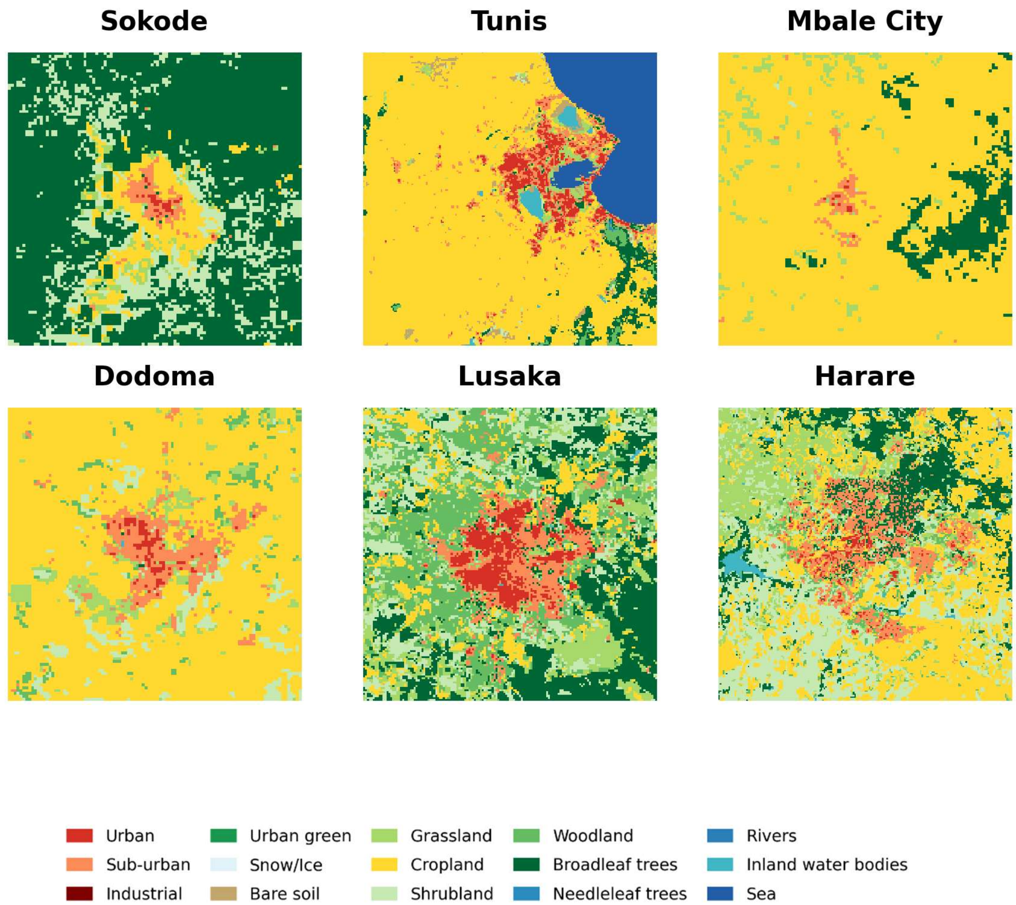
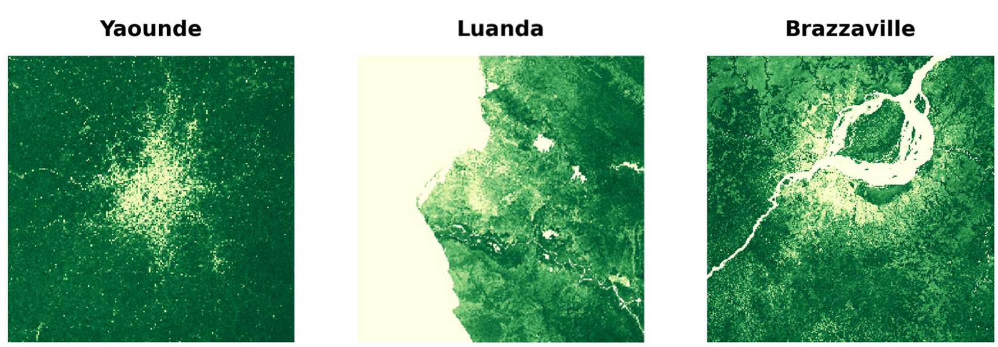


Figure 11 UrbClim input Land Cover of all the 54 CAIAC cities.

Vegetation fraction

We only show here a few examples of UrbClim vegetation fraction input:



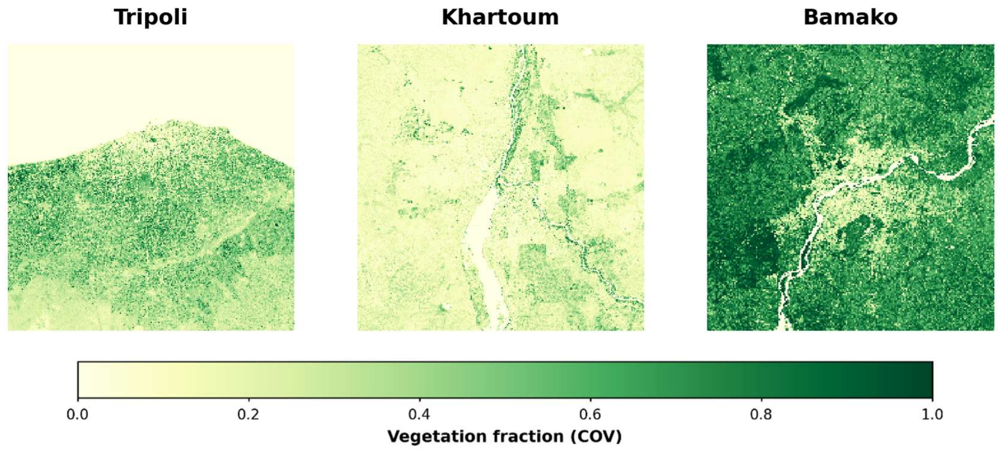


Figure 12 UrbClim vegetation fraction input for Yaounde, Luanda, Brazzaville, Tripoli, Khartoum and Bamako