LATE REQUEST FOR A EMI R&D PROJECT 2026–2028

Country/Organisation: Italy / CIMA Research Foundation

Principal Investigator[1]: Andrea Zonato

Affiliation: CIMA Research Foundation

Address: CIMA Research Foundation, Savona, Italy
andrea.zonato@cimafoundation.org

Other researchers: Luca Monaco; Massimo Milelli

Project Title: **GLIDE-SOL and DL-SOL for pan-European urban thermal-stress forecasting, ensembles, and probabilistic UTCI services**

To make changes to an existing project please submit an amended version of the original form.

If this is a continuation of an existing project, please state the computer project account assigned previously.		
Starting year: (A project can have a duration of up to 3 years, agreed at the beginning of the project.)	2026	
Would you accept support for 1 year only, if necessary?		X

Late EMI R&D Projects are normally approved for 1 year, but exceptionally can be approved for 3 years.

Computer resources required for project year:		2026	2027	2028
High Performance Computing Facility	[SBU]	8000000	12000000	18000000
Graphics Processing Unit Cluster-A	[GBU]	750000	1500000	2500000
Graphics Processing Unit Cluster-B	[GBU]	250000	500000	500000
Accumulated data storage (total archive volume)[2]	[GB]	30000	80000	140000

[1] The Principal Investigator will act as contact person for this EMI R&D Project and, in particular, will be asked to register the project, provide annual progress reports of the project's activities, etc.
[2] These figures refer to data archived in ECFS and MARS. If e.g. you archive x GB in year one and y GB in year two and don't delete anything you need to request x + y GB for the second project year etc.
[3]The number of vGPU is referred to the equivalent number of virtualized vGPUs with 8GB memory.

EWC resources required for project year:		2026	2027	2028
Number of vCPUs	[#]	256	512	512
Total memory	[GB]	1024	1536	2048
Storage	[GB]	15000	20000	30000
Number of vGPUs[3]	[#]	8	16	16

Continue overleaf.

Principal Investigator: Andrea Zonato

Project Title: GLIDE-SOL / DL-SOL pan-European UTCI services

Extended abstract

All EMI R&D Project requests should provide an abstract/project description including a scientific plan, a justification of the computer resources requested and the technical characteristics of the code to be used. The completed form should be submitted/uploaded at https://www.ecmwf.int/en/research/special-projects/special-project-application/special-project-request-submission.

Following submission by the relevant Member State the EMI R&D Project requests will be published on the ECMWF website and evaluated by ECMWF and its Scientific Advisory Committee. The requests are evaluated based on their scientific and technical quality, and the justification of the resources requested. Previous EMI R&D Project reports and the use of ECMWF software and data infrastructure will also be considered in the evaluation process.

Requests exceeding 10,000,000 SBU should be more detailed (3-5 pages).

Project summary

This late EMI R&D request seeks phased CPU and GPU support to scale GLIDE-SOL and its deep-learning surrogate DL-SOL on ECMWF infrastructures in a credible research-to-service progression. GLIDE-SOL is an end-to-end urban thermal-stress forecasting framework that converts numerical weather prediction and high-resolution geospatial city data into hourly UTCI maps. It couples a turnkey preprocessor for urban morphology and forcing preparation with a GPU-accelerated implementation of SOLWEIG, thereby resolving mean radiant temperature, shadows, sky-view factors, wind-related urban modifiers, and downstream thermal-stress indicators at meter-scale resolution. The main scientific goal is to move from current single-city and limited multi-city demonstrations toward a benchmarked pan-European prototype capable of deterministic forecasts, targeted ensembles, retrospective reanalyses, and probabilistic UTCI products.

Scientific objectives for 2026-2028

The project has four tightly connected objectives. First, we will deploy operational high-resolution UTCI forecasting for a large European city set using harmonized urban inputs and regular numerical weather forcing, starting from current ICON-EU-based workflows and extending them to ECMWF-compatible forecast and reanalysis environments. Second, we will develop DL-SOL, a surrogate model trained on GLIDE-SOL outputs, to emulate hourly Tmrt fields while preserving the physical structure learned from the original radiation solver. Third, we will perform systematic sensitivity and feature-importance experiments to quantify the relative contribution of morphology, vegetation, land cover, terrain, wind exposure, and meteorological forcing to urban thermal variability. Fourth, we will produce deterministic, ensemble, and final probabilistic radiant-load and UTCI products suitable for rapid-risk assessment, service prototyping, and retrospective climatological analyses.

Planned yearly work

Year 1 will focus on production hardening of GLIDE-SOL on ECMWF resources, city onboarding, automation of preprocessing, and explicit benchmark campaigns to document CPU, GPU, and storage demand against existing reanalysis-driven experiments. Year 2 will expand the spatial portfolio, scale the ensemble workflow in a controlled way, and train DL-SOL on a cross-city archive derived from the physically based system. Year 3 will emphasize probabilistic output generation, extended hindcast datasets, cross-climate transfer testing, and the generation of research-grade evaluation products for publication and future service transition. The request is therefore intentionally phased: the first year establishes validated baselines, the second year supports scale-up where benchmarks justify it, and the third year extends only the components that have demonstrated scientific and technical value.

Why both CPU and GPU resources are required

The codebase is not a pure GPU kernel. The workflow has a substantial CPU side that includes large geospatial downloads, reprojection, raster harmonization, land-use stacking, building and wall processing, forcing assembly, quality control, job orchestration, compression, and post-processing of hourly rasters and NetCDF outputs. These tasks require parallel CPU throughput, memory, and I/O bandwidth. The GPU side is equally central because the radiative solver becomes the dominant cost once multiple cities, forecast cycles, or ensemble members are executed. The request is therefore balanced but deliberately restrained: CPU resources support preprocessing, archive generation, and high-volume post-processing, while GPU resources are centred on Cluster-A as the main execution target, with limited Cluster-B usage retained only for portability tests and queue-resilience fallback.

Technical characteristics of the code

GLIDE-SOL is implemented in Python and relies on PyTorch for accelerated kernels and DL-SOL development, plus xarray, netCDF4, numpy, scipy, rasterio, geopandas, shapely, pyproj, pandas, and related scientific I/O libraries. The code already supports multi-GPU execution, tile decomposition, and distributed synchronization. The preprocessor automatically builds consistent input packages including DEM, building DSM, walls, aspect, vegetation, land use, optional wind coefficient layers, and hourly forcing files. The runtime can write large collections of GeoTIFF and NetCDF products and is therefore both compute-intensive and I/O-intensive. DL-SOL will add a second class of GPU workloads: repeated training, validation, cross-city transfer tests, and inference over large spatial grids. Together, these characteristics justify using both conventional HPC and GPU partitions rather than relying on a single resource type.

Resource justification

Our current reference-city baseline already demonstrates a substantial demand, but it does not justify asking immediately for full operational-service scale. A typical complete workload consists of six independent one-day SOLWEIG cases per city, each requesting about four GPUs, sixteen CPU cores, one hundred twenty-eight gigabytes of memory, and roughly three hours of wall time, corresponding to about eighteen 4-GPU node-hours per city before including sensitivity tests, alternative forcings, ensemble members, DL-SOL training data generation, or reruns for quality control. Our current ECMWF project itusrrm1 is allocated 7,000,000 SBU per year and has already used about 91 percent of that budget, which supports the case for growth but also provides a concrete baseline for a phased request. For that reason, the Year-1 request is kept at a moderate R&D scale, with larger Year-2 and Year-3 envelopes to be activated only if benchmarking confirms the need. The requested CPU SBU supports preprocessing and archive-generation campaigns, while the GPU GBU primarily supports the SOLWEIG kernel, DL-SOL training, and a controlled expansion of ensemble experiments.

Storage and EWC justification

Storage growth remains significant, but the archive request has been scaled to a research-credible level. A full city package, including final outputs and the main reproducibility layers, can approach about one hundred gigabytes at current settings, and larger archives are expected once we retain selected ensemble outputs, probabilistic summaries, and the training corpora required by DL-SOL. The accumulated archive therefore grows from 30 TB in 2026 to 140 TB in 2028, rather than assuming operational retention volumes from the outset. EWC resources are requested only for orchestration services, pre- and post-processing APIs, staging of harmonized urban inputs, lightweight interactive products, and rapid DL-SOL inference. The vCPU, memory, storage, and vGPU envelopes have been reduced to a modest R&D profile consistent with prototype services rather than permanent large-scale operations.

Requested resource scenarios

Scenario	CPU / GPU envelope	Archive storage	EWC profile	Use case
Minimum	2026 only: 5M SBU, 0.5M GBU	30 TB	128-256 vCPU, up to 1 TB RAM, 10-15 TB, 4-8 vGPU	Benchmarking, limited city set, baseline DL-SOL experiments
Recommended	2026-2027: 8-12M SBU, 1-2M GBU	30-80 TB	256-512 vCPU, 1-1.5 TB RAM, 15-20 TB, 8-16 vGPU	Controlled multi-city scaling and ensemble prototyping
Full scale	Growth toward 18M SBU and 3M GBU by 2028	Up to 140 TB	Up to 512 vCPU, 2 TB RAM, 30 TB, 16 vGPU	Only after benchmark-documented expansion and validated need

Expected outcome and relevance to ECMWF

The main outcome will be a scalable platform for near-real-time urban heat monitoring and forecast-based thermal-risk assessment across Europe, together with a surrogate model that reduces turnaround time enough to support broader ensemble and probabilistic use. Scientifically, the project will clarify the drivers of urban thermal exposure and quantify uncertainty in city-scale radiant-load predictions. Operationally, it will provide a directly usable research prototype for hourly UTCI mapping, rapid what-if experimentation, and the generation of spatially explicit thermal-risk products from ECMWF-compatible meteorological forcing. If the full multi-year trajectory cannot be granted, the work can still proceed under the minimum or recommended scenario by reducing the number of simultaneously active cities, the ensemble depth, and the annual retraining frequency of DL-SOL.