REQUEST FOR AN EMI R&D PROJECT 2027–2029

Country/Organisation:	Spain
Principal Investigator[1]:	PhD Researcher Arnau Toledano Rubí (atoledano@meteo.ub.edu)
Affiliation:	Universitat de Barcelona, Departament de Física Aplicada - Meteorologia
Address:	Martí i Franquès, 1. 08028 Barcelona. Floor 7
Other researchers:	Dra. Mireia Udina Sistach (mudina@meteo.ub.edu), Dra. Yolanda Sola Salvatierra (ysola@meteo.ub.edu), Dr. Joan Bech Rustullet (joan.bech@ub.edu).
Project Title:	*Enhancing Numerical Weather Prediction in Complex Terrain by Integrating State-of-the-Art AI Techniques*

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.	**SP** …………..	
Starting year: (A project can have a duration of up to 3 years, agreed at the beginning of the project.)	2027	
Would you accept support for 1 year only, if necessary?	YES X	NO ☐

Computer resources required for project year:		**2027**	**2028**	**2029**
High Performance Computing Facility	[SBU]	45,000,000	70,000,000	50,000,000
Graphics Processing Unit Cluster-A	[GBU]	4,000,000	5,000,000	4,000,000
Graphics Processing Unit Cluster-B	[GBU]	4,500,000	5,500,000	4,500,000
Accumulated data storage (total archive volume)[2]	[GB]	250,000	400,000	400,000

[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:		**2027**	**2028**	**2029**
Number of vCPUs	[#]	64	70	64
Total memory	[GB]	320	350	320
Storage	[GB]	20,000	25,000	20,000
Number of vGPUs[3]	[#]	2	2	2

Principal Investigator: Arnau Toledano Rubí (atoledano@meteo.ub.edu)

Project Title: *Enhancing Numerical Weather Prediction in Complex Terrain by Integrating State-of-the-Art AI Techniques*

Extended abstract

1. Background and scientific motivation

Numerical weather prediction (NWP) over complex terrain remains one of the most demanding problems in operational and research meteorology. Many of the physical processes that govern local weather — orographic flows, thermally driven circulations, boundary-layer turbulence — occur at scales below the resolution of operational mesoscale models and must therefore be parameterised. The Weather Research and Forecasting model (WRF-ARW), one of the most widely used limited-area NWP systems in the scientific community and operational services, solves the compressible non-hydrostatic Euler equations on a structured grid. A major limitation is that its computational cost increases steeply as horizontal resolution increases, owing to the growing number of grid cells and the shorter timestep required for numerical stability.

It is well established that mesoscale and microscale NWP exhibit systematic biases in complex terrain, arising primarily from the smoothed representation of topography and from subgrid-scale parameterisations, particularly those governing the atmospheric surface layer. In parallel, AI-based global forecasting systems (Pangu-Weather, GraphCast, AIFS) have demonstrated that, once trained, data-driven models can match or exceed traditional NWP systems such as ECMWF's IFS for several variables and lead times, at a fraction of the inference cost. However, their extension to local and complex-terrain scales remains largely unexplored: representing fine-scale topography and meso/microscale physics is an open problem that this project addresses directly.

This work is carried out within a PhD project funded by the FI-STEP predoctoral grant (STEP 2025 00137) at the Universitat de Barcelona, and builds on the group's long-standing observational and modelling programme in the Pyrenees (ATMOUNT, Cerdanya-2017 campaign).

2. Objectives

General objectives

1. Develop a hybrid AI–NWP system to improve weather prediction in complex terrain (e.g. the Pyrenees).
2. Advance the physical understanding of surface-layer processes through ML-assisted discovery of empirical relationships from simulation and observational data.

Specific objectives (research lines)

- **L1 — AI-based meteorological downscaling.** Development of deep learning architectures for downscaling of surface meteorological variables, progressing from deterministic models (attention U-Net architectures) to probabilistic and generative approaches (latent diffusion models) for uncertainty quantification.
- **L2 — WRF-LES simulation of complex terrain.** Generation of high-resolution reference datasets via WRF coupled with Large-Eddy Simulation (WRF-LES) over selected case studies, validated against surface observation networks and field campaigns.
- **L3 — Surface-layer parameterisation correction.** Development of grey-box ML correction schemes (progressing through Lasso, LightGBM, MLP, and symbolic regression with PySR) targeting known failure modes of Monin–Obukhov Similarity Theory (MOST)-based surface schemes (MM5/Revised, MYNN) in complex terrain, with the corrected scheme implemented and tested online as a new physics module within WRF.

3. Methodology and relation to the requested project years

The PhD work plan spans November 2025 to October 2028 across five tasks. The EMI R&D Project request (2027–2029) covers the computationally core:

- **2027 — Control simulation completion and generative downscaling.** Completion of the ERA5- and, where available, ERA6-driven WRF control-simulation dataset across the nested hierarchy (9 / 3 / 1 / 0.333 / 0.111 km) for the selected case studies, and the start of training for the probabilistic/generative downscaling architectures (L1, second phase). This combines WRF-LES integration at the two finest nests with early-stage multi-GPU training.
- **2028 — Generative model training and online surface-layer coupling.** Full training campaign for the diffusion/latent-diffusion downscaling models, together with the development of the grey-box surface-layer correction (L3): this stage requires iterative WRF re-compilation and re-execution to test the corrected parameterization online against the LES reference, which is the single most resource-intensive phase of the project.
- **2029 — Validation, sensitivity testing, and closure.** Completion of L3 verification runs, sensitivity experiments in response to peer review of the resulting possible manuscripts, and any additional case studies required to close out the dataset. Compute demand decreases relative to 2028 but remains elevated due to the need for replication and testing.

This progression is reflected in the requested HPCF and GPU allocation across the three years.

4. Technical characteristics of the code

Dynamical core / NWP component

- WRF-ARW v4.6.1, run in real-data mode with initial/boundary conditions from ERA5 reanalysis (1-hourly). ERA6, once available for the relevant study periods, and operational HRES-IFS data, subject to applicable data-access arrangements, will also be considered.
- Nested domain configuration: 9 km → 3 km → 1 km → 0.333 km → 0.111 km, two-way nesting.
- Physics suite: Thompson microphysics, RRTMG longwave/shortwave radiation, MYNN PBL/surface-layer scheme (baseline, subject to the L3 correction), Noah-MP land-surface model.
- Vertical resolution: more than 90 vertically stretched levels, with enhanced resolution near the surface for the LES nests.
- Given the cost of the finest nests, simulations are restricted to selected case studies / type days relevant to each research line (e.g. orographic windstorms, anticyclonic thermal inversion episodes) rather than continuous integration.

AI / deep learning component

- Framework: PyTorch, trained on GPU-accelerated partitions with mixed-precision and multi-GPU data-parallel training.
- L1: attention U-Net for deterministic downscaling; latent diffusion / generative architectures for probabilistic downscaling and uncertainty quantification.
- L3: grey-box pipeline (Lasso → LightGBM → MLP → PySR symbolic regression) using MOST-derived inputs (Obukhov length, friction velocity, surface roughness parameters) to correct similarity functions; final scheme implemented as a Fortran physics module coupled into WRF for online testing.

5. Computational resource justification

Resource estimates are based on measured scaling benchmarks from the group's existing WRF experiments on the Universitat de Barcelona's infrastructure. They were extrapolated using core-hour consumption, domain grid-point counts, timestep constraints, and the expected numbers of control, LES-reference, sensitivity, and online-coupling experiments. A technical pilot at the start of 2027 will refine per-case estimates and optimise domain decomposition, I/O strategy and GPU data pipelines.

(i) WRF-LES / HPCF (SBU). Cost scales primarily with grid-point count (∝ $1/dx^2$), vertical levels, number of timesteps (∝ 1/dx for stability under the same Courant number), and the number of case studies simulated per year. The finest nests (0.333 km and 0.111 km) dominate total cost despite covering a smaller area, since cost per nest grows roughly as dx^{-3} for a fixed physical domain and fixed-CFL timestep scaling.

Year	Main HPCF Workload	Requested HPCF [SBU]
2027	Completion of WRF/WRF–LES control dataset, selected 0.333/0.111 km simulations, benchmark sensitivity experiments, preprocessing and verification.	45,000,000
2028	Repeated online WRF testing of candidate surface-layer corrections against LES reference simulations; expanded sensitivity matrix and replication across meteorological regimes.	70,000,000
2029	Independent validation cases, final physics-sensitivity experiments, replication, robustness testing and reviewer-driven additional simulations	50,000,000

The requested CPU allocation is concentrated in 2028 because the online parameterization-correction phase requires repeated re-compilation, re-execution and evaluation of WRF under a common set of chosen complex-terrain cases. Unlike offline ML correction, each candidate formulation must be evaluated through full dynamical feedback within WRF and against LES-scale reference simulations.

(ii) Deep learning training / GPU Cluster-A and Cluster-B (GBU). GPU resources are crucial for the probabilistic and generative components of L1 and for the iterative ML development in L3. The requested allocation is distributed across the two available GPU infrastructures because the project will use them for complementary workloads. Cluster-A (NVIDIA A100) will be used for reproducible baseline experiments, deterministic and probabilistic U-Net variants, ablation studies, hyperparameter searches and medium-scale training. Cluster-B (Grace Hopper / H100) will prioritise the largest mixed-precision distributed training runs, including latent-diffusion models, high-resolution spatio-temporal architectures, ensemble generation and long autoregressive or multi-step evaluation experiments.

Year	Workload	Requested GPU-A [GBU]	Requested GPU-B [GBU]
2027	Deterministic and probabilistic baseline models; initial latent-diffusion experiments; dataset and pipeline optimisation	4,000,000	4,500,000
2028	Full generative-model campaign; multi-GPU diffusion training; L3 grey-box retraining and online-coupling iterations	5,000,000	5,500,000
2029	Independent-event validation; uncertainty calibration; replication; sensitivity experiments and final model consolidations	4,000,000	4,500,000

(iii) Storage. Archive volume is driven primarily by the WRF/WRF-LES output (netCDF4, 3D fields at high temporal frequency for the finest nests) and by trained-model checkpoints. Estimated cumulative archive volume, calculated according to the EMI accounting rules, is 250,000 GB (2027), 400,000 GB (2028), 400,000 GB (2029, partial cleanup of intermediate LES output after dataset consolidation), all within the 736 TB per-project cap.

6. Data management and use of ECMWF infrastructure

ERA5 reanalysis already serves as the initial/boundary-condition source for the group's WRF simulations, so this would extend an existing dependency on ECMWF data infrastructure rather than introduce a new one. Observational validation draws on SMC and AEMET surface networks and the ATMOUNT/Cerdanya-2017 field campaign. This is the PI's first EMI R&D Project request; no prior EMI/Special Project account exists.

Model output, trained-model checkpoints, and processing code will be retained on departmental storage during the active research phase, with restricted access to the project team, and published via Zenodo and the CSUC CORA data repository upon completion of each associated article, following Universitat de Barcelona CRAI data-management guidelines.

The **European Weather Cloud** will be used as the project’s active data-staging, preprocessing and reproducibility environment. It will host subsets of WRF/WRF–LES output and observational data, metadata catalogues, preprocessing workflows, interactive quality-control and visualisation scripts/notebooks, and lightweight inference and verification services. The EWC will support efficient movement of curated datasets and model artefacts between ECMWF archive services, the HPCF and the Universitat de Barcelona infrastructure. Full WRF integration and large-scale distributed deep-learning training will remain on the HPCF and GPU clusters; EWC vGPU resources will be used for development, pipeline validation and lightweight inference rather than primary model training.

7. Expected outcomes

The project is expected to produce 3–4 peer-reviewed articles in meteorology/atmospheric science and AI venues, supporting the PI PhD thesis defence, targeted for 2029. Results have direct relevance to operational forecasting of hazards in complex terrain (heavy snowfall, strong orographically forced winds, and pollution episodes associated with thermal inversions) and the corrected surface-layer scheme is intended to be transferable to other NWP systems using MOST-based parameterizations.

Acknowledgement

Acknowledgement is made for the use of ECMWF's computing and archive facilities in this research.