REQUEST FOR AN EMI R&D PROJECT 2027–2029

Country/Organisation: Switzerland

Principal Investigator[1]: Nora Zilibotti

Affiliation: ETH Zurich

Address: Universitätstrasse 16, 8006 Zürich, Switzerland

Other researchers: Emmanouil Flaounas (Hellenic Centre for Marine Research, Greece)
Prof. Heini Wernli (ETH Zurich, Switzerland)

Project Title: AI-Driven Online Correction of Model-Error Growth Across Scales and Domains in NWP

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 ☒	NO ☐

Computer resources required for project year:		**2027**	**2028**	**2029**
High Performance Computing Facility	[SBU]	10,000,000	15,000,000	
Graphics Processing Unit Cluster-A	[GBU]			
Graphics Processing Unit Cluster-B	[GBU]	3,000,000	4,000,000	
Accumulated data storage (total archive volume)[2]	[GB]	25,000	50,000	

[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	[#]			
Total memory	[GB]			
Storage	[GB]			
Number of vGPUs[3]	[#]			

Continue overleaf.

Principal Investigator: Nora Zilibotti

Project Title: AI-Driven Online Correction of Model-Error Growth Across Scales and Domains in NWP

Extended abstract

1. Scientific background and motivation

Despite decades of progress in numerical weather prediction (NWP), deterministic forecast skill remains limited to lead times of about ten days (Zhang et al., 2019). A substantial part of this limitation arises not only from initial-condition uncertainty, but also from inherent model error mainly linked to the approximate representation of parametrized processes that are unresolved, or only partially resolved, by the prognostic equations, such as turbulence, convection, radiative transfer and cloud microphysics (Bauer et al., 2015; Palmer, 2017). These errors are generated locally, but interact, amplify and eventually upscale during the forecast, so that the simulated atmospheric state diverges progressively from reality often within few forecast days, degrading the skill of key variables such as wind, temperature and precipitation (Sánchez et al., 2020; Judt, 2018).

In contrast to the extensive research approach, developed to reduce uncertainties in initial conditions, there is currently no systematic, online method that directly targets the growth of model errors during a forecast simulation. In fact, model errors are only addressed indirectly through increased horizontal resolution, improved parameterisations, and long development-and-validation cycles. These approaches translate into operational gains but do not act on the atmospheric conditions through which model error actually grows in time. This is a particularly acute limitation for applications, such as wind-energy forecasting, where for medium-range wind forecasts mean absolute percentage errors typically reach 15–20% and can exceed 80% in high-wind conditions (Olsen et al., 2025).

This project tests a hybrid physical–AI forecasting system in which an AI module is trained to learn the atmospheric conditions under which model error grows and how to correct these errors from past forecasts. Once trained, the AI module acts directly on the resolved model state while the numerical model performs a forecast to suppress that growth. The module will be designed to emulate the corrective tendencies of spectral nudging without requiring knowledge of a reference large-scale state (e.g. reanalysis), and therefore imposes no intrinsic limit on spatial resolution nor is restricted to already known atmospheric states. This comes in contrast to data-driven emulators bounded by reanalysis resolution and by their training distribution.

2. Objectives

The primary objective of the project is the development of a prototype AI module that reproduces the spectral-nudging correction at coarse resolution over the Euro-Atlantic domain.

Secondary objectives involve the transferability of the AI module to other forecast domains and resolutions, with a focus on the US and Europe. In this context, horizontal resolution and forecast lead times will be gradually increased to assess transferability to industrial applications.

3. Scientific plan and methodology

3.1 Phase 1 — Proof of concept

The Weather Research and Forecasting model (WRF) is used as a testbed. A large ensemble of 7-day forecasts is performed over the Euro-Atlantic domain at a coarse horizontal resolution of ~30 km. Both initial and boundary conditions are provided by ERA5 for all simulations, so that initial conditions can be treated as "perfect" and any forecast bias attributed to the production, growth and upscale propagation of model error; the coarse resolution also makes WRF output directly comparable to ERA5, against which a "perfect forecast" is defined.

In this phase, spectral nudging is applied at meso- to large scales (≈1,000 km). Spectral nudging acts like an additional parameterisation, supplying temperature and momentum tendencies whose sole purpose is to prevent the model-resolved state from diverging from a reference field (ERA5) — i.e. to suppress the growth of model error. Preliminary tests show excellent WRF–ERA5 agreement under spectral nudging even at 7-day lead times. Because spectral nudging requires the reference state, it can only correct past forecasts and cannot be used in real time. For every nudged simulation we archive, as standard model output, the resolved atmospheric fields, the tendencies produced by all WRF parameterisations, and the additional tendencies applied by spectral nudging. Together these constitute the training dataset.

An AI module is then trained to reproduce the nudging tendencies aiming to supress growth of errors — as spectral nudging would, but without access to the reference large-scale state. Architectures of increasing complexity are evaluated, from pointwise feedforward networks (testing whether local information suffices) to 3D convolutional / U-net architectures operating on limited X×Y×Z subvolumes of a limited number of variables. Subvolume sizes range from about 20×20×5 grid cells in coarse resolutions to about 100×100×20 in higher ones, the larger receptive field being needed to capture the synoptic context from which model-error growth upscales. Predictor sets, kernel sizes and hyperparameters will be optimised systematically. Successful development should be demonstrated through skillful forecasts with the AI module approaching the performance of a simulation where spectral nudging is applied.

3.2 Phase 2 — Transferability and scalability

The twin experiments are repeated at progressively higher resolutions (20, 10 and 5 km) than the 30 km at which the module was trained, quantifying the extent to which forecast error is intrinsic to the parameterisations rather than dependent on resolved scales, and hence whether retraining is required when the configuration changes. In parallel we determine the minimum training-set size needed to recover optimal performance, and conduct limited-area very-high-resolution (~1 km) retraining experiments, testing the retrained module on domains different from those used for training (including a US domain). This phase delivers guidelines for cost-efficient training while retaining skill, and a prototype module transferable to high-resolution operational configurations.

4. Technical components of the project

Atmospheric model: WRF (Weather Research and Forecasting model, ARW core, v4.7), compiled for the Atos HPCF and run as a distributed-memory application across multiple nodes. Spectral nudging is used through WRF's native capability. Initial and boundary conditions are taken from ERA5 (retrieved from MARS); standard and custom diagnostic output (resolved fields plus parameterisation and nudging tendencies) is written for training-data generation.

AI module: implemented in Python, comprising feedforward and 3D convolutional / U-net networks/operators. Training is performed offline on GPU using WRF outputs. For the twin experiments the trained AI module will be coupled to WRF for online inference, evaluated within the time-stepping loop so that the corrective tendencies are applied at every model step over the forecast integration.

5. Justification of the requested computing resources

The request spans two distinct facilities: the HPC for all WRF integrations, and the GPU clusters for training the AI module. Data will be archived in ECFS/MARS.

5.1 High-Performance Computing Facility resources

A single 7-day WRF forecast over the Euro-Atlantic domain at 30 km costs of order $1.2–2.5×10^4$ SBU, depending on node count and I/O. 2027 (Phase 1): generation of the training ensemble (several hundred nudged 7-day forecasts spanning seasons and regimes) together with the proof-of-concept twin experiments accounts for the requested 8 million SBU. 2028 (Phase 2): the higher-resolution twin experiments (20/10/5 km) and limited-area ~1 km nests are substantially more expensive per simulation — cost scales steeply with resolution — so a smaller number of integrations, plus cross-domain simulations, accounts for the requested SBUs.

5.2 GPU resources

At this stage a precise GPU requirement cannot be determined, because two quantities that drive the cost are subject to research results: (1) the size of the training dataset which depends on how many nudged forecasts prove necessary to span the relevant atmospheric regimes and (2) the complexity of the AI architecture, e.g. the number of layers and channels and the patch volume of the 3D U-net, which will be selected during the project. We therefore provide an order-of-magnitude estimate intended to cover the expected range rather than a precise figure.

Training cost is dominated not by single simulations but by the architecture and hyperparameter search and by Phase-2 retraining. As a working estimate, a representative full training of a 3D U-net on patch volumes in the stated range costs of order a few tens of GPU-hours (taken here as ≈30 GPU-hours per training). The architecture screening, predictor and hyperparameter search, and the Phase-2 retraining across resolutions and domains together require of order 50–100 such trainings, giving a programme total of roughly 1,500–3,000 GPU-hours. On this basis we request approximately 3 million GBU in 2027 (~830 GPU-hours, early screening on the Phase-1 dataset) and 4 million GBU in 2028 (≈1,100 GPU-hours, full search and Phase-2 retraining), about 7 million GBU in total. Because the estimate is necessarily approximate, actual usage will be monitored and reported in the annual progress report, and refined against the reserve if required.

5.3 Data storage

The training dataset — resolved fields, all parameterisation tendencies, and spectral-nudging tendencies, output at sub-daily frequency for several hundred forecasts — dominates storage. We request 25 TB in 2027 and 50 TB cumulative in 2028 (Phase-2 higher-resolution output adds to, rather than replaces, the Phase-1 archive; ECMWF archive volume is billed cumulatively). Both figures are well within the per-project ceiling.

Table 1. Summary of requested resources.

Resource	2027	2028
High-Performance Computing Facility [million SBU]	8.0	12.0
GPU Cluster-B – GH200 [million GBU]	3.0	4.0
Accumulated data storage [TB]	25	50

References

Bauer P., Thorpe A., Brunet G. (2015) The quiet revolution of numerical weather prediction. Nature, 525, 47–55.

Judt F. (2018) Insights into atmospheric predictability through global convection-permitting model simulations. J. Atmos. Sci., 75, 1477–1497.

Olsen B.T.E. et al. (2025) Low-level jets in the North and Baltic seas: mesoscale model sensitivity and climatology using WRF V4.2.1. Geosci. Model Dev., 18, 4499–4533.

Palmer T. (2017) The primacy of doubt: evolution of NWP from determinism to probability. J. Adv. Model. Earth Syst., 9, 730–734.

Sánchez C., Methven J., Gray S., Cullen M. (2020) Linking rapid forecast error growth to diabatic processes. Q. J. R. Meteorol. Soc., 146, 3548–3569.

Zhang F. et al. (2019) What is the predictability limit of midlatitude weather? J. Atmos. Sci., 76, 1077–1091.