REQUEST FOR AN EMI R&D PROJECT 2027–2029

Country/Organisation: Sweden

Principal Investigator[1]**:** Daniel Yazgi

Affiliation: Swedish Meteorological and Hydrological Institute (SMHI)

Address: Folkborgsvägen 17, 601 76, Norrköping, Sweden

Other researchers:
Fuxing Wang (SMHI)
Swapan Mallick (SMHI)

Project Title:

Multiple Meteorological Applications Using ML-Based Generative Diffusion Models

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.)	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]	2 M		
Graphics Processing Unit Cluster-A	[GBU]	7.2M		
Graphics Processing Unit Cluster-B	[GBU]	6.8M		
Accumulated data storage (total archive volume)[2]	[GB]	10,000 GB		

[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: Daniel Yazgi

Project Title: Multiple Meteorological Applications Using ML-Based Generative Diffusion Models

Extended abstract

This proposal presents three AI-driven research projects in weather and climate science. Utilizing state-of-the-art generative diffusion models, these projects leverage a unified codebase and require high-performance GPU resources. Brief overviews of each project are detailed below.

Project 1

The new generation climate models, so called Convection-Permitting Models (CPMs), provide a step-change in climate simulations by explicitly resolving convection at kilometer-scale resolutions. CPMs run at km-scale spatial resolution and capture fine-scale surface heterogeneities, thereby improving the representation of extremes (Prein et al., 2015). However, their high computational cost limits production of multi-decadal, multi-scenario and large ensemble simulations (Kendon et al., 2021).

This project aims to apply generative diffusion machine learning models (see Projects 2 and 3 for more descriptions) to emulate Harmonie-Climate (HCLIM; Wang, 2024) regional climate model simulations at convection-permitting scale (3 km resolution) from coarse resolution inputs (12km resolution). First, we will set up the diffusion model for the Fenno-Scandinavia domain using the NorCP simulations by HCLIM (Lind et al., 2020; Wang et al, 2024). Second, we will evaluate model performance over both historical and future climate scenarios. Third, we will generate ensembles of climate simulations under various future scenarios and evaluate their performance.

The project covers two geographic regions (northern Italy and Fenno-Scandinavia) with varying grid dimensions to test model transferability. Each domain includes at least two driving Global Circulation Models (GCMs) and at least two climate change scenarios, covering historical and future periods (mid-century and end-of-century). The storage is estimated to be around 4,000 GB to 5,000 GB. Based on our previous applications, we estimate our minimum resource requirement to be 5,000,000 GBUs.

Project 2

Recent progress in next-generation artificial intelligence (AI)-based diffusion models has revealed significant potential for uncertainty quantification in numerical weather prediction and climate science. Diffusion models offer a robust machine learning framework capable of capturing intricate uncertainty relationships and reconstructing flow-dependent uncertainty structures from limited datasets. In our recent study, we developed, evaluated, and validated a novel diffusion-based machine learning system designed to learn the uncertainty relationship between the ERA5 Ensemble of Data Assimilations (ERA5-EDA) and the high-resolution Copernicus Arctic Regional Reanalysis version 2 (CARRA2, Mallick et al, 2026). The proposed framework produces high-resolution gridded uncertainty fields and exhibits a strong ability to capture flow-dependent uncertainties during independent validation periods. The results further indicate that the model effectively identifies uncertainty patterns linked to complex orographic features, thereby demonstrating its capacity to represent realistic spatial variability over heterogeneous terrain. Additionally, the diffusion model successfully reconstructs both physically consistent uncertainty fields and the corresponding high-resolution physical state variables, yielding outputs that closely approximate the reference CARRA2 fields. These outcomes underscore the robustness, reliability, and generalizability of diffusion-based approaches for generating spatially coherent and physically plausible uncertainty estimates. The method is the denoising diffusion probabilistic model (available at https://github.com/CARRA2/Uncertainty_Quantification).

Building on these encouraging findings, the proposed research will explore the integration of diffusion-based methodologies with the AI-driven WeatherGenerator (www.weathergenerator.eu) forecasting framework and the high-resolution CARRA2 dataset at a 2.5-km spatial resolution. This integration is anticipated to substantially enhance the representation of local-scale atmospheric processes, cloud microphysics, and severe weather events, while concurrently improving uncertainty-aware high-resolution forecasting capabilities. Moreover, the project will investigate hybrid physics–AI approaches, advanced conditioning techniques, and probabilistic observation operators to develop a comprehensive end-to-end AI-driven data assimilation framework including conventional and remote sensing data assimilation. This integrated system aims to synergize data-driven learning with physical process understanding, thereby advancing next-generation probabilistic weather forecasting and uncertainty quantification methodologies.

A major challenge of the previous work was the substantial computational cost associated with diffusion models and the limited availability of computational resources. All previous experiments were performed on the ECMWF ATOS supercomputing system under restricted resource allocations, where a maximum of five GPU-enabled nodes per user was available, each containing four GPUs. Model training and evaluation were typically conducted using four GPUs on a single node, often requiring long queue times. Training a model for a single variable, up to 20,000 iterations required approximately 10 hours. The evaluation and validation stages were even more computationally demanding because the processing time increased with the number of generated samples used for uncertainty quantification and field reconstruction. For instance, generating 20 samples from each of 10 trained models required more than six hours of computation for a single analysis time. The previous project utilized approximately 216,000 GBUs in total. The proposed extension of this work to data assimilation systems will require a similar, if not larger, allocation of computing resources to enable model development, extensive training, sensitivity experiments, and large-scale ensemble generation. In addition, approximately 2 million CPU hours will be required for machine-learning training data preparation, preprocessing, post-processing, and validation activities.

Project 3

Deep learning is now widely used in weather applications because it is faster and cheaper to run than physical models (Tian et. al. 2026). Running diffusion models (DMs) directly on the full atmospheric state is expensive in memory and sampling time, on top of a high training cost (Rombach et. al. 2022). To benefit from DMs without that cost, the meteorological fields are compressed into a compact latent space with a variational autoencoder, and further tasks are carried out on this latent representation. We used such a representation to interpolate in time between two atmospheric states. The main difficulty is making the latent representation accurate enough, particularly to represent extreme events such as heavy precipitation.

This work builds on a time-interpolation project that uses a diffusion model to regenerate atmospheric states between frames separated by six hours. The first goal is to improve the existing interpolation model which is based on EDM (Karras e. Al. 2022) and test it on a larger domain of about 1000x1000, since it is currently tested only on 256x256. The second goal is to produce hourly forecasts instead of interpolation with uncertainty quantification using the diffusion model. The third goal is to apply the same diffusion model on the climate data for time interpolation. The dataset is ~550GB and the needed storage for operation is

~150GB which in total is ~700GB. The computational capacity needed is about 350000 GBUs full training. Which makes the maximum 14M GBU needed to do multiple training and fine tunings.

References

Karras, T., Aittala, M., Aila, T. & Laine, S. Elucidating the Design Space of Diffusion-Based Generative Models. Preprint at https://doi.org/10.48550/arXiv.2206.00364 (2022).

Kendon E. J., A. F. Prein, C. A. Senior, A. Stirling, 2021: Challenges and outlook for convection-permitting climate modelling. *Phil. Trans. R. Soc. A.,* **379**(2195), 20190547, https://doi.org/10.1098/rsta.2019.0547.

Mallick, S., Bojarova, J., Schyberg, H., Dahlgren, P., Støylen, E., Yang, X., and Tølløse, K. Copernicus pan-Arctic Regional Reanalysis (CARRA2): Uncertainty estimation method. ECMWF Report, 2026. Available at: https://confluence.ecmwf.int/display/CKB/Copernicus+pan-Arctic+Regional+Reanalysis+%28CARRA2%29%3A+Uncertainty+estimation+method

Prein, A. F., and Coauthors, 2015: A review on regional convection-permitting climate modeling: Demonstrations, prospects, and challenges. *Rev. Geophys.*, **53**, 323–361, https://doi.org/10.1002/2014RG000475.

P. Lind, D. Belušić, O.B. Christensen, A. Dobler, E. Kjellström, O. Landgren, D. Lindstedt, D. Matte, R.A. Pedersen, E. Toivonen, F. Wang. Benefits and added value of convection-permitting climate modeling over Fenno-Scandinavia Clim. Dyn., 55 (7) (2020), pp. 1893-1912, 10.1007/s00382-020-05359-3

Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B. High-Resolution Image Synthesis with Latent Diffusion Models. Preprint at https://doi.org/10.48550/arXiv.2112.10752 (2022).

Tian, J. et al. ARROW: An Adaptive Rollout and Routing Method for Global Weather Forecasting. Preprint at https://doi.org/10.48550/arXiv.2510.09734 (2026).

Wang, F. (2024). An introduction to the HARMONIE-Climate (HCLIM) regional climate modeling system. Zenodo. https://doi.org/10.5281/zenodo.11424Fuxing, 181

Wang, F., Landgren, O., Laakso, A., Christensen, O. B., Dobler, A., & Lind, P. (2024). User policy for NorCP data. Zenodo. https://doi.org/10.5281/zenodo.13827561