HAMLET-PHYSICS 2024, Copenhagen, 19-21 August 2024

Data-driven modelling for limited area forecasting

Simon Adamov (MCH), Leif Denby (DMI), Tomas Landelius (SMHI), Fredrik Lindsten (LiU), Joel Oskarsson (LiU), Thomas Rieutord (Met Eireann), Irene Schicker (GeoSphere Austria), Michiel Van Ginderachter (RMI)

Introduction

Recent work by Oskarsson et al. 2023 [1] has demonstrated with *neural-lam* that it is possible to train a graph neural network to produce purely data-driven km-scale weather forecasts in a restricted spatial domain using the traditional Limited Area Modelling (LAM) approach - as used by operational centres across the world to produce sub-daily high-resolution weather forecasts.

Based around the open-source *neural-lam* code base [2] the ML LAM Community [3] has formed to with the aim to further develop this modelling capability. We here report on our work since forming and the outlook for our community as a whole.

Work so far has principally focused around:

1. Surveying and building an inventory of regional km-scale datasets that could be used to train on.
2. Developing common tools to convert these datasets from the source file formats and structure (often GRIB or netCDF) and to a common format (zarr), to create training datasets
3. Developing common tools for constructing the neural network graph
4. Developing common data-stores interface to allow ingesting many different data formats
5. Carrying out scientific investigation to analyse how to best to couple the limited area model to the global model forecast.

Datasets survey and preparation

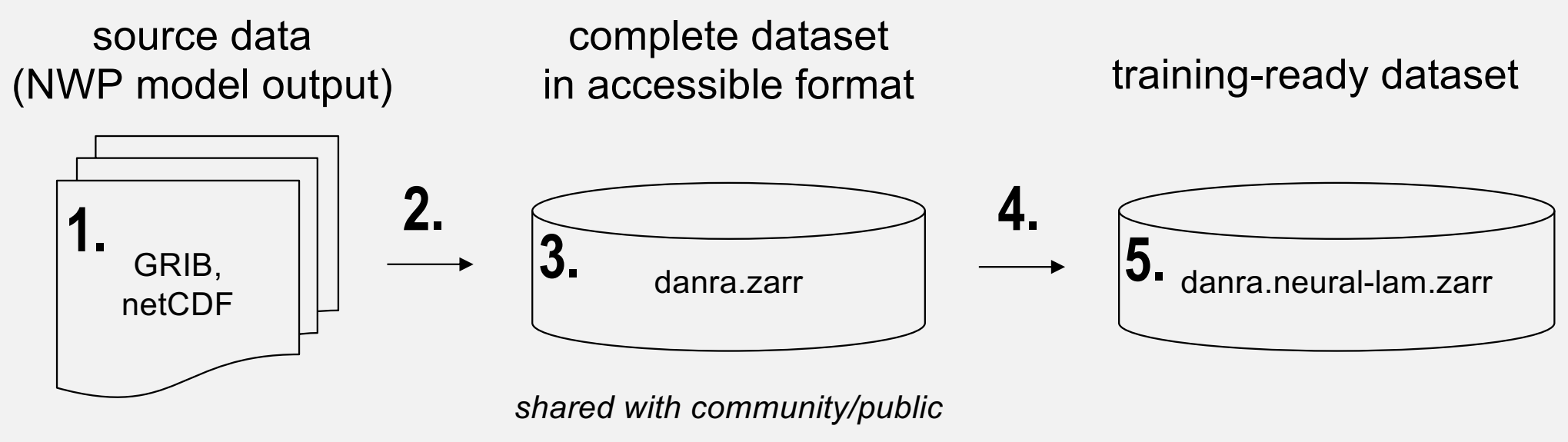
What datasets already exist?

Below we have reproduced the current dataset survey carried out within the community. It is the aim that all these datasets will be converted to a common zarr-based format and shared openly within the community.

DANRA reanalysis	2.5km, 3-hourly	30 years	Northern Europe centred on Denmark	GRIB	Complete GRIB collection processed to zarr stored at DMI, being transferred to EWC (1-year sample already on EWC)
MEPS	10km (can get 2.5km), 1-hourly	~2 years		GRIB, numpy-arrays	Internally at SMHI + numpy-arrays at Berzelius (LiU). Contains 5 ensemble members.
MÉRA	2.5km	35 years	British Isles	GRIB1	Currently stored on ECFS + local copy at Met Éireann https://www.met.ie/climate/available-data/mera
ARA	2.5 km, 1-hourly	5 - 10 years		GRIB or netcdf	Currently being simulated, once finished it will be on data.hub.geosphere.at
CERRA reanalysis	5.5 km	~35 years	Europe	GRIB/netcdf	Subset on Copernicus Data Store, full fields on EFS (ECMWF)
CARRA reanalysis	2.5km	~30 years	Artic		CDS: https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-carra-model-levels
NORA3 Hindcast	3 km	50 year 1970-2023	Slanted across Bristish Isles and Scandinavia	netCDF	https://thredds.met.no/thredds/projects/nora3.html,
NEWA macro-runs	3 km	210 years	Europe	netcdf	New European Wind Atlas Access via API requests, seems like web access is restricted now. Limited in parameters and vertical levels (targeted wind energy)
DOWA	2.5 km	10 years 2009-2019	The Netherlands	netcdf	Dutch Offshore Wind Atlas https://www.dutchoffshorewindatlas.nl/about-the-atlas/dowa-data/data-downloads https://www.wins50.nl/data/
COSMO-2 (Operational Analysis based on KENDA)	2.2km	5 years	Switzerland + surroundings (Alps)	GRIB2 / netCDF / Zarr	MeteoSwiss will open-source data between 2024-2026 (new law)
COSMO-REA2 reanalysis	2km	7 years 2007-2013	Central Europe	?	https://reanalysis.meteo.uni-bonn.de/?Download_Data___COSMO-REA2
WINS50 reanalysis	2.5km	3 years 2019-2021 (DOWA extension)			https://api.dataplatform.knmi.nl/open-data/v1/datasets/wins50_wfp_nl_daily_3d/versions/1/files and ec:/nkl/harmonie/DOWA/DOWA_40h12tg2_fERA5/

Why and how are we processing datasets to zarr?

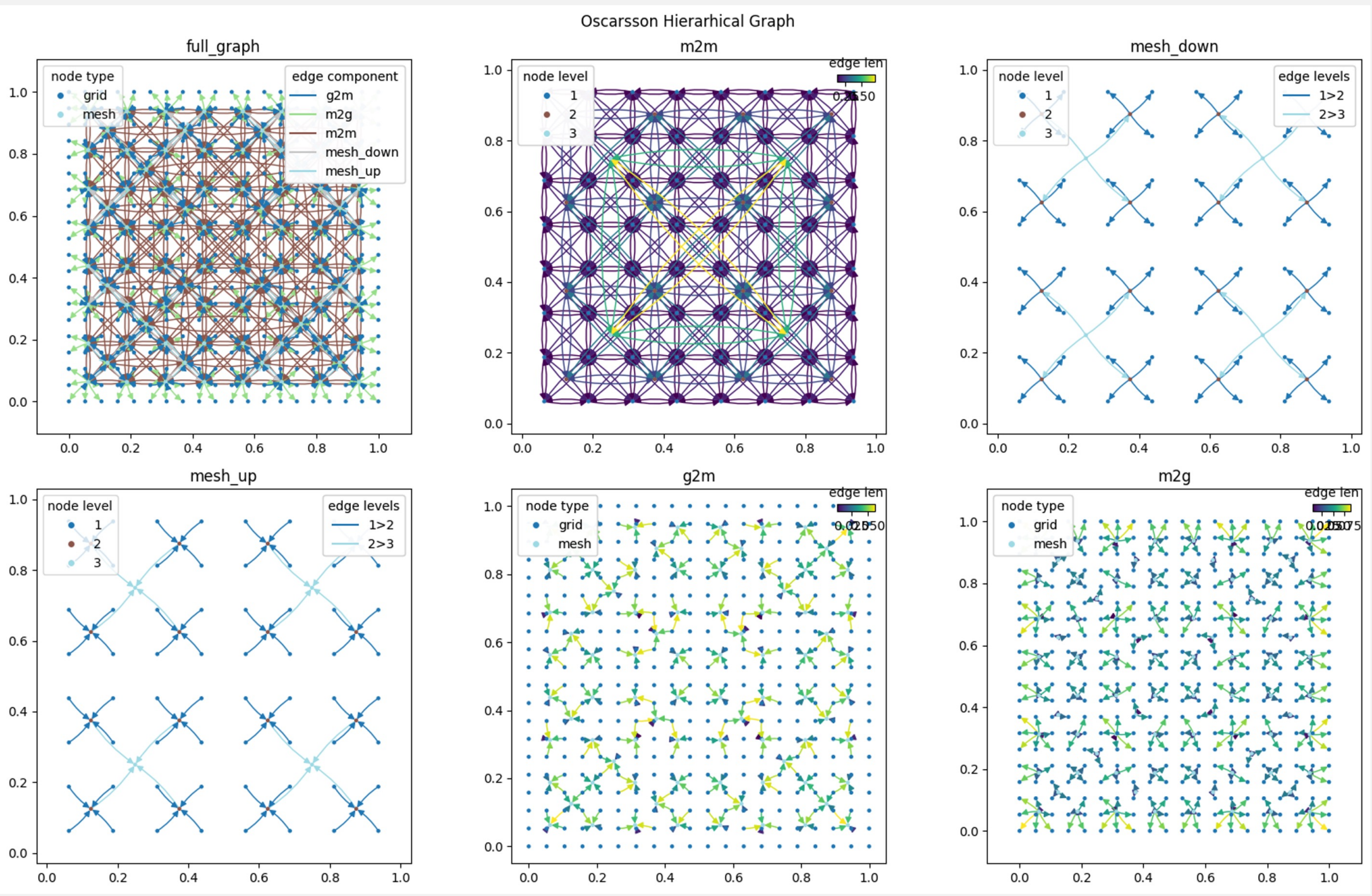
Why: zarr enables fast, easy and efficient retrieval of any subset (time/space) of any variable in dataset. This is because in zarr data-values of individual variables and coordinates are stored in separate file chunks, each chunk spanning a fixed block of the data, with the variables, coordinates and meta-data stored in a predefined file-structure. Functionally, zarr appears like netCDF to the end-user (supported natively by *xarray* python package and built into NETCDF-C library).


source data (NWP model output)
complete dataset in accessible format
training-ready dataset
1. GRIB, netCDF
2.
3. danra.zarr
4.
5. danra.neural-lam.zarr
shared with community/public

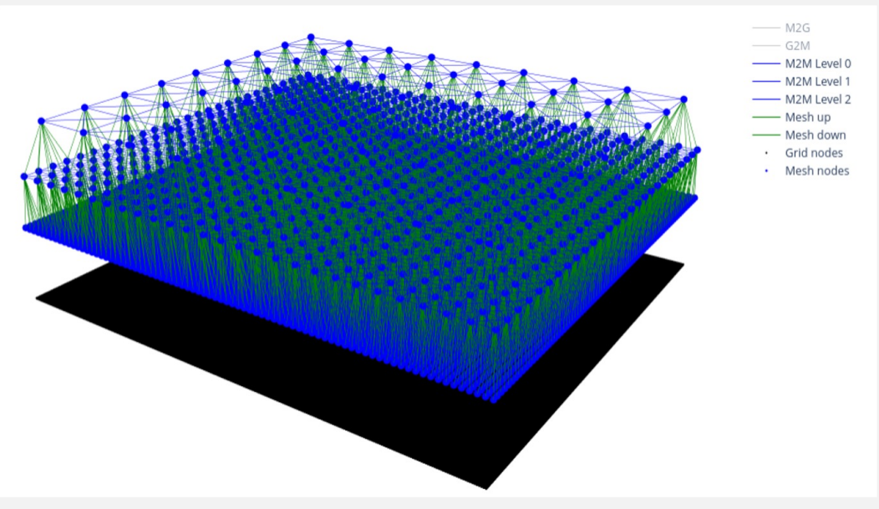
1. Source data (reanalysis, reforecasts) in various formats and file/directory structure
2. Transformation of all source files into a single (or collection of) zarr-based datasets
    a. Tools: danra_to_zarr (https://github.com/leifdenby/dmi-danra-to-zarr), gribscan-harmonie (https://github.com/leifdenby/gribscan-harmonie)
3. Complete zarr-based dataset containing all data across space and time
4. Transformation to training-ready dataset
    a. *create* in anemoi-datasets (https://github.com/ecmwf/anemoi-datasets), mllam-data-prep (https://github.com/mllam/mllam-data-prep)
5. Training-ready zarr-based dataset targeting specific model architecture

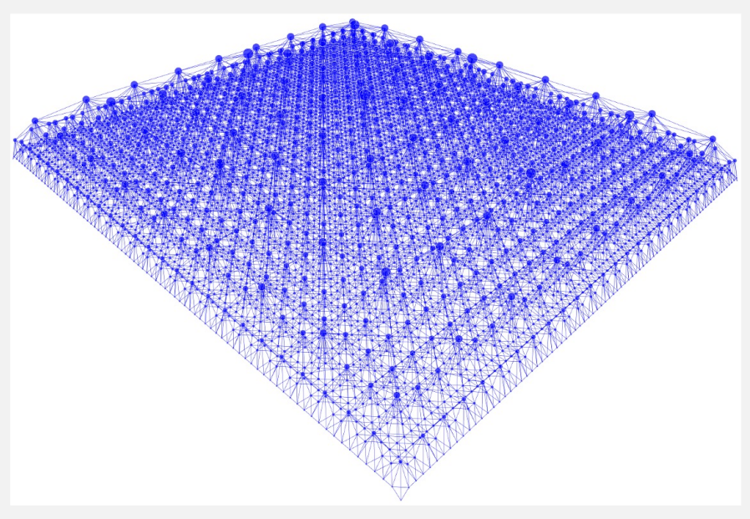
Tooling for constructing the graph

To aid the further development of different graph architectures in *neural-lam* and graph-based weather models in general, functionality has been developed within *neural-lam* to construct the graph used with the graph neural-network by creating individual graph components. This separates the creation of the graph components (represented as *networkx.DiGraph* objects) that comprise the different parts of the message-passing [4] graph; grid2mesh (encode), mesh2mesh (process) and mesh2grid (decode), from the serialisation into *pytorch_geometric.Data* datastructures which are loaded into the model. The separation of this latter step enables targeting different of code-bases that implement graph-based weather forecasting modelling.



This tooling also comprises customisable visualisation which enables development of documentation and learning material (Jupyter notebooks) and interactive creation of new exotic graph architectures.



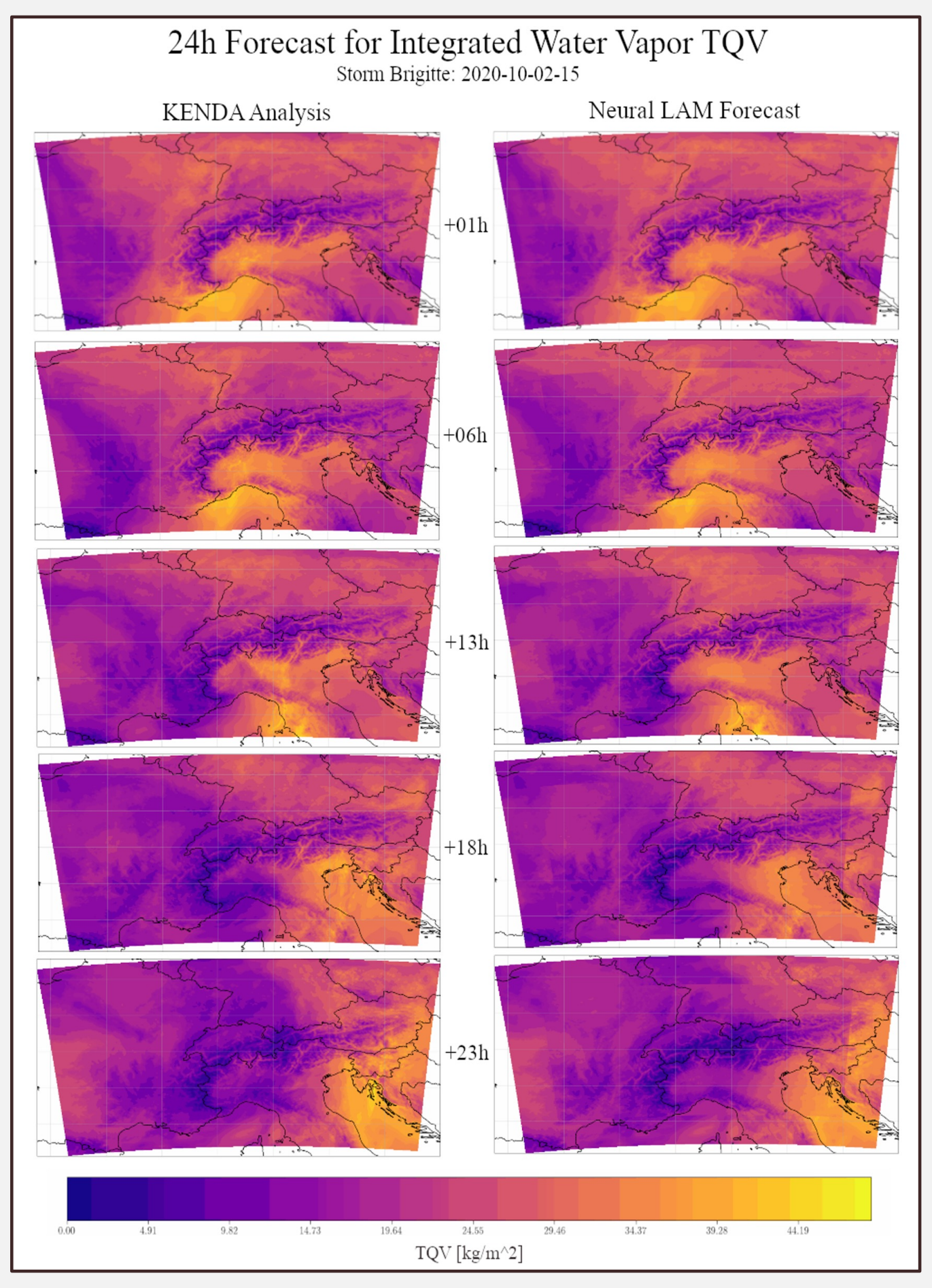


Roadmap:

- Integrate more performant graph-object container data-structure (e.g. pyg.HeteroData, graph_tool.Graph)
- Add functionality to construct hexagonal and triangular mesh graphs

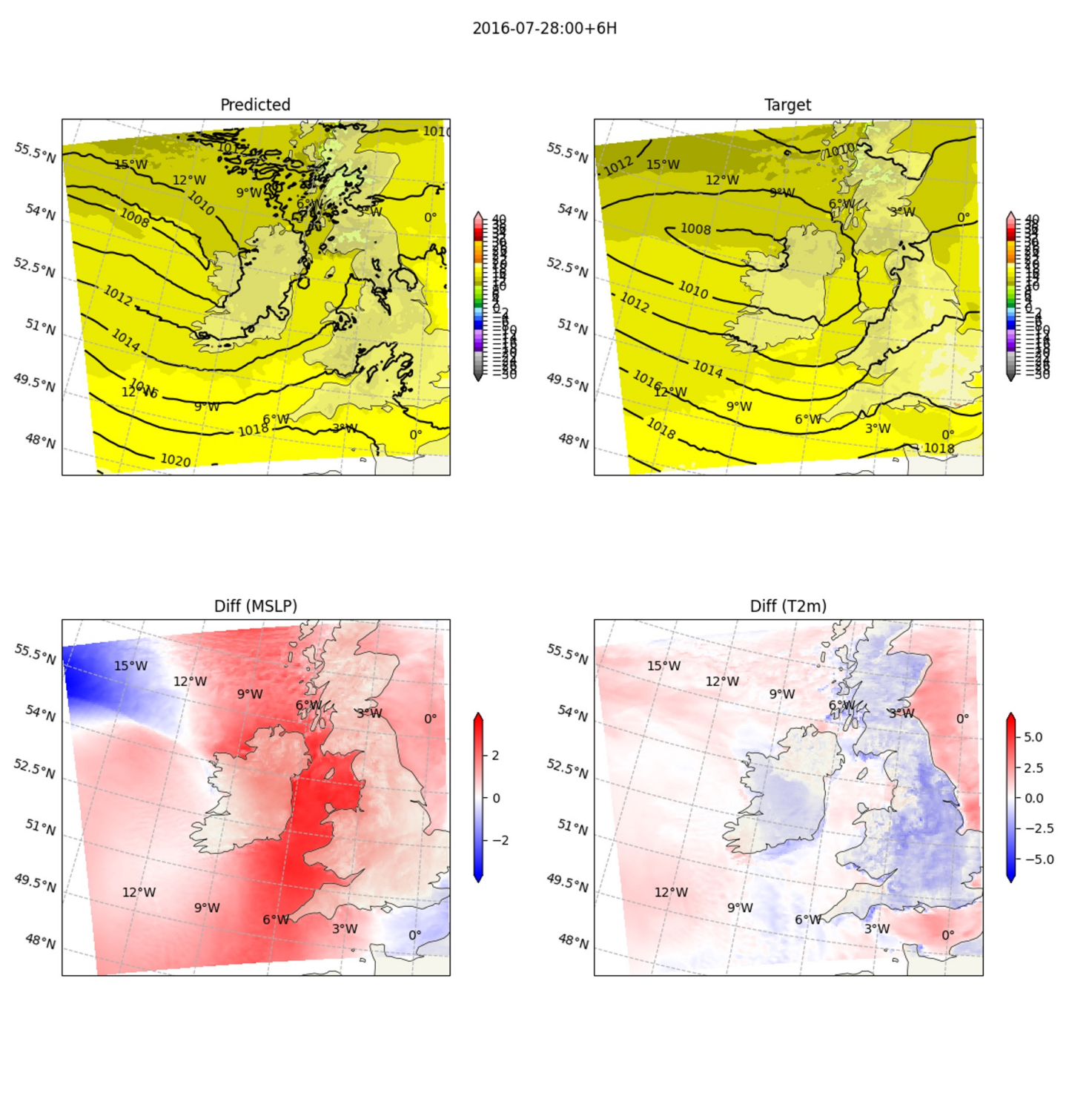
Updates on modelling efforts

Alpine Domain (MeteoSwiss)



Trained on 4 years of hourly COSMO-2.2 km operational analysis data (KENDA), Neural-LAM generates 24h forecasts for a broad range of surface and free atmosphere variables. The graph-based architecture performs well over steep topography. The plot on the left showcases such an autoregressive forecast for integrated water vapor (TQV). During storm Brigitte significant water vapor parcels enter the domain from a 60 km boundary (30 outer grid cells).

Irish domain (Met Éireann)



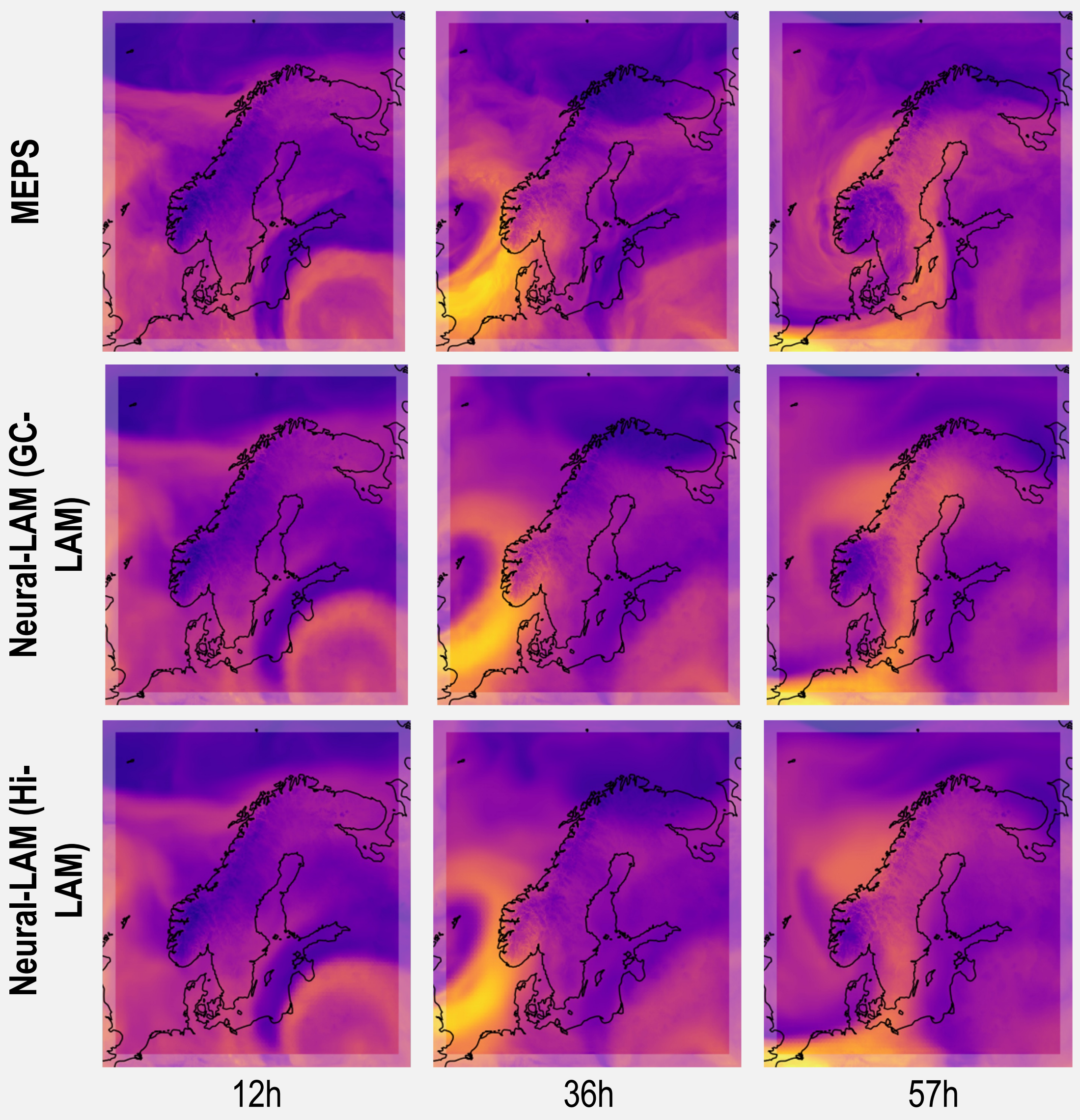
Trained on 1 year of data from the MÉRA reanalysis with the same settings as in [1] (GC-LAM). Isolines are MSLP, colors are T2m. Despite limited training, the output fields are realistic. However, they show wrinkles that grow bigger with lead time and advection is not captured properly.

Updates on modelling efforts

Nordic Domain [1]

- Using data from MetCoOp Ensemble Prediction System (MEPS), Neural-LAM is used to build a fast surrogate mode.
- Trained on a dataset containing 10 forecasts per day from a period of 1 year.
- Single model for forecasting 17 variables (surface + atmosphere) up to lead time 57h.
- Experiments both with multiscale graph (GC-LAM), similar to GraphCast [5], and new hierarchical graph construction (Hi-LAM).

Example forecast: Integrated water vapor



Ongoing Efforts

- Probabilistic modeling for ensemble forecasting using variation autoencoders [6]
- Realistic boundary conditions using ERA5 and IFS
- Integration of NeuralLAM with the Anemoi Framework of ECMWF, by utilisation of common tools (for data-preparation, graph-generation and prediction validation) and by integration of model architectures and other developments into Anemoi.
- Perturbation of latent space (weights in graph mesh) for ensemble augmentation and xAI
- Perturbation of static fields
- Integration into existing verification pipelines to validate against observations (gridded and point), looking into specific metrics for selected applications (for example SPEI for wind-energy applications).
- Transfer learning. Can a model trained on one region be transferred to another one or at least provide added value in terms of reduced training time or amount of training data?
- Building common data catalog of zarr-based processed datasets that can be used to create training datasets

References

[1]: Graph-based Neural Weather Prediction for Limited Area Modeling, Oskarsson et al. 2023, NeurIPS 2023 Workshop on Tackling Climate Change with Machine Learning 2023, https://arxiv.org/abs/2309.17370

[2]: original repo: https://github.com/joeloskarsson/neural-lam, community fork: https://github.com/mllam/neural-lam

[3]: The ML LAM Community meets monthly (everyone is welcome!), details, meeting notes and development plan on https://bit.ly/mllam-plan

[4]: Relational inductive biases, deep learning, and graph networks, Battaglia et al 2018, https://arxiv.org/abs/1806.01261

[5]: Learning skillful medium-range global weather forecasting, Lam et al. 2023, Science

[6]: Probabilistic Weather Forecasting with Hierarchical Graph Neural Networks, Oskarsson et al. 2024, https://arxiv.org/abs/2406.04759