MAchinE Learning for Scalable meTeoROlogy and climate

Report on the application of ML solutions within the W&C workflow

www.maelstrom-eurohpc.eu

MAELSTROM

D1.6 Report on the application of ML solutions within the W&C workflow

Author(s):	Thomas Nipen (MetNor) and all application developers of Work Package 1
Dissemination Level:	Public
Date:	31/03/2024
Version:	0.1
Contractual Delivery Date:	31/03/2024
Work Package/ Task:	WP1/T1.5
Document Owner:	MetNor
Contributors:	4cast, ECMWF, ETH, FZJ
Status:	Final

MAELSTROM

Machine Learning for Scalable Meteorology and Climate

Research and Innovation Action (RIA)
H2020-JTI-EuroHPC-2019-1: Towards Extreme Scale Technologies and Applications

Project Coordinator:	Dr Peter Dueben (ECMWF)
Project Start Date:	01/04/2021
Project Duration:	36 months

Published by the MAELSTROM Consortium

Contact:
ECMWF, Shinfield Park, Reading, RG2 9AX, United Kingdom
Peter.Dueben@ecmwf.int

The MAELSTROM project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 955513. The JU receives support from the European Union’s Horizon 2020 research and innovation programme and United Kingdom, Germany, Italy, Luxembourg, Switzerland, Norway

Contents

Figures

1 Executive Summary

This deliverable describes the work done on the six MAELSTROM applications to either operationalize them or make them production-ready. This deliverable also describes how the MAELSTROM workflow tools, the MAELSTROM benchmark datasets, and knowledge generated from the project have been integrated into the workflow at the institutes involved in the project.

By the end of the project, MAELSTROM delivered one application running operationally (AP1), one application demonstrated within an operational environment (AP3), and four production-ready applications (AP2, AP4, AP5, and AP6) ready to be exploited. MAELSTROM also contributed to the development of ai-models, a tool used operationally to run several emerging data-driven weather models (AP8).

2 Introduction

2.1 About MAELSTROM

To develop Europe’s computer architecture of the future, MAELSTROM will co-design bespoke compute system designs for optimal application performance and energy efficiency, a software framework to optimise usability and training efficiency for machine learning at scale, and large-scale machine learning applications for the domain of weather and climate science.

The MAELSTROM compute system designs will benchmark the applications across a range of computing systems regarding energy consumption, time-to-solution, numerical precision and solution accuracy. Customised compute systems will be designed that are optimised for application needs to strengthen Europe’s high-performance computing portfolio and to pull recent hardware developments, driven by general machine learning applications, toward needs of weather and climate applications.

The MAELSTROM software framework will enable scientists to apply and compare machine learning tools and libraries efficiently across a wide range of computer systems. A user interface will link application developers with compute system designers, and automated benchmarking and error detection of machine learning solutions will be performed during the development phase. Tools will be published as open source.

The MAELSTROM machine learning applications will cover all important components of the workflow of weather and climate predictions including the processing of observations, the assimilation of observations to generate initial and reference conditions, model simulations, as well as post-processing of model data and the development of forecast products. For each application, benchmark datasets with up to 10 terabytes of data will be published online for training and machine learning tool-developments at the scale of the fastest supercomputers in the world. MAELSTROM machine learning solutions will serve as blueprint for a wide range of machine learning applications on supercomputers in the future.

2.2 Scope of this deliverable

2.2.1 Objectives and work performed in this deliverable

The objective of this deliverable is to document the integration of the MAELSTROM applications, workflow tools, and associated knowledge into the workflow at the partner institutes. This work was carried out in Task 1.5.

The work performed on each of the six MAELSTROM applications is described in Section 3. This includes any further scientific work that was done to improve the ML-solutions since D1.4 and any work to operationalize the applications. We also describe the work done to integrate and further exploit the MAELSTROM workflow tools, datasets, and knowledge at each of the respective institutes.

Although not part of the MAELSTROM original proposal, we describe in Section 3.7 how MAELSTROM has contributed to ECMWF’s efforts in developing the Artificial Intelligence/Integrated Forecasting System (AIFS), a fully data-driven weather prediction model.

A summary of the work is found in Section 4.

2.2.2 Deviations and countermeasures

There are no deviations to the work plan in this deliverable.

3 Application integration

3.1 AP1: Blend citizen observations and numerical weather forecasts

3.1.1 Background

In application one, we developed a deep-learning model to post-process temperature forecasts from a high-resolution Numerical Weather Prediction (NWP) model for the Nordics.

Throughout the MAELSTROM project, we developed a U-Net model with 6 levels (see D1.4 for details on the architecture). The machine learning model takes inputs from the MetCoOp Ensemble Prediction System (MEPS), a 2.5 km resolution limited area model, and includes fields such as temperature, precipitation, winds, and clouds. Static topographical parameters, such as land area fraction and altitude are also included (both for the NWP model grid and target grid, which are at different resolutions). The target data for the application is a high-resolution gridded analysis (1 km) based heavily on citizen observations from Netatmo's network of personal weather stations. These target datasets were also used as input to the model (in the form of model biases over the 24 hours leading up to the forecast initialization). The output of the model is probabilistic and represented by 3 quantile fields (10th, 50th, and 90th percentiles) of the forecast distribution. The loss function for this application was the quantile loss.

The end goal of this application was a product of high enough quality to be suitable for operational use at MET Norway. This model was integrated into production February 27, 2024, replacing a very simple linear regression scheme using a small subset of the predictors that was operationalized in 2018 (Nipen et al. 2020). These forecasts are used on MET Norway's public weather app Yr (https://www.yr.no).

3.1.2 Further developments since D1.4

A number of scientific improvements have been made to the ML solution since D1.4.

Shuffling of data. Firstly, the shuffling of input data during training was significantly improved. The previous implementation processed all 58 timesteps from a single forecast run before reading the next. Due to limits on available memory on the nodes, only a few files could be kept in memory at a time and be shuffled. This reduced our ability to create batches with randomised samples. Instead, we changed the data loader to only read single timesteps from each forecast run. As the timesteps in the files are stored sequentially without chunking, retrieving a single timestep did not incur a significant IO penalty. This allowed us to shuffle across dates with very different climates. Shuffling significantly improved training results (Fig. 1a). The overall loss was reduced, and the training was a lot more stable.

Assessing the value of added complexity. We also assessed the underlying assumption that deep-learning leads to improvements for this application. We tested if a more complex model led to higher forecast accuracy. This was done by training U-Nets with different numbers of levels. A 1-level U-Net is equivalent to a simple convolutional neural network. As the number of levels increases, the number of trainable parameters increases. A 1-level U-Net has 611 trainable parameters, a 2-level U-Net has 4,579, a 3-level U-Net has 20,195, and a 6-level U-Net has 1,314,019. We found that the validation loss improved monotonically with an increasing number of levels (Fig. 1b), confirming that the architecture is able to extract deeper signals with more complexity.

Assessing the value of big data. We also assessed whether or not a longer training period improved out-of-sample forecast accuracy. Our full training dataset contained 362 forecast days spanning a 1-year period. We decreased this to 24 days, by selecting 2 days from each month of the year. As expected, a larger training set led to a lower overall validation loss. With a small dataset, overfitting starts being noticeable after only 9 epochs (Fig. 1c). These last two results provide evidence that deep-learning models combined with large datasets are advantageous for post-processing.

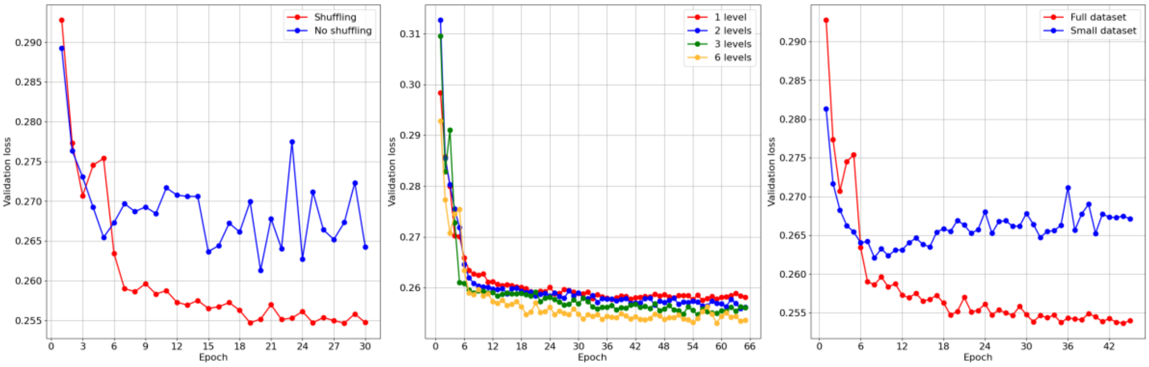


Fig. 1: Validation loss as a function of epoch, showing (a) the importance of shuffling the data, (b) different number of U-Net levels, and (c) the importance of using the whole training dataset.

Weather-dependent correction. One of the motivations for using an advanced post-processing method is its assumed ability to correct the NWP model for weather-dependent biases. Previous efforts at MET Norway have used bias-correction methods that are not dependent on the weather situation. Given a measure of biases over the last 24 hours, we have traditionally corrected the bias forward in time based only on the forecast lead time. The weights for these corrections were optimised across all weather patterns and we have long suspected that this is suboptimal. We were therefore curious to check if the MAELSTROM-developed method can correct the input NWP forecast differently based on the weather. We did this by feeding the network synthetic data with different weather and measuring the increment that the model predicts. We set the biases of the NWP model for the previous 24 hours to 2°C and studied the correction predicted by the model for different lead times under different weather situations. The two weather dependent effects we are interested in are: 1) Do biases persist differently during precipitation events vs dry events? 2) Do cold biases persist differently than warm biases?

We found that the ML model does indeed predict that the persistence of bias is weather-dependent (Fig. 2). Firstly, biases are moderated much more strongly under precipitation events than under dry conditions. This is expected, because under precipitation events, the atmosphere is generally more mixed and biases in high-resolution models are less common. Secondly, the model allows cold biases to persist longer into the future than warm biases. We are unsure why this is the case. Both examples show evidence of the mechanisms behind why deep learning models can provide better forecasts than simple models that do not allow for weather-dependent biases.

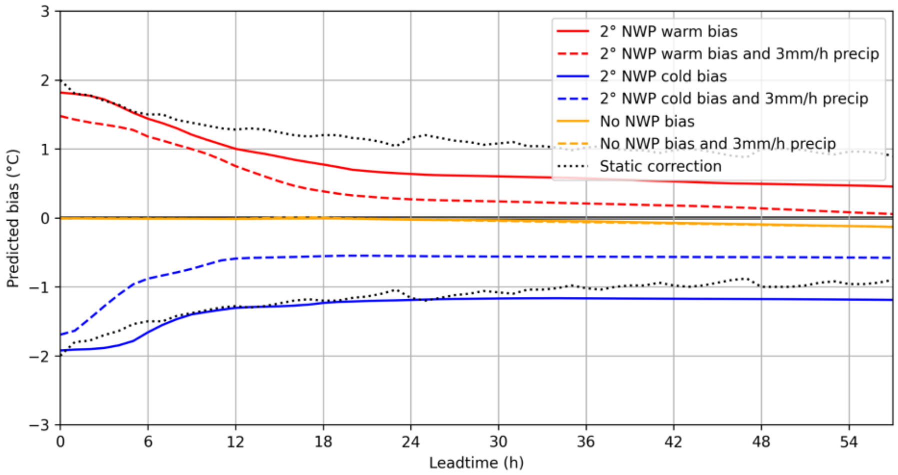


Fig. 2: The ML-model's prediction of the NWP bias as a function of lead time given that the NWP model biases over the last 24 hours are +2°C (red), -2°C (blue), and 0°C (orange). Dashed lines show cases where 3 mm/h of precipitation is present. The dotted black lines show the static bias-correction used operationally at MET Norway prior to the introduction of this new method.

3.1.3 Operational implementation

The U-Net implementation and data pipeline developed in MAELSTROM was implemented into MET Norway's operational post-processing system for weather forecasting. This system produces weather forecasts for the Nordics every hour, using the most up-to-date observations and NWP information available.

Our operational post-processing chain typically takes 22 minutes from the time the NWP models and observations are ready to when our data is available to our end-users on Yr. As we do not have dedicated GPU nodes for operational use, we run the U-Net model inference on CPUs. This adds around 2 minutes to our processing time compared to before, which is not ideal, but acceptable for our use case.

These forecasts are now part of our fully automated forecast chain, which does not have any manual approval process or other form of human intervention in place before the forecasts reach our end users. We have therefore done extensive evaluation work to make sure that the forecasts can be trusted.

3.1.4 Further exploitation of results at MET Norway

The workflow tools and experience gained from MAELSTROM has already shaped our machine learning strategy for operational post-processing at MET Norway. We rely extensively on post-processing to bridge the gap between the output from NWP models and the high resolution localised information our users expect. Our vision is that the model and data pipelines developed in MAELSTROM will form the backbone of our post-processing infrastructure for years to come, and that we will use this for improving other parameters in MET Norway's portfolio of products. To generate our public weather forecasts, we process data for 3 domains (Nordic, Arctic, global) and 4 time-scales (nowcast, short-range, medium-range, subseasonal-range). In total, we post-process close to 100 weather parameters from a wide range of NWP models at different resolutions. Finding robust ways to handle this in a scalable way is important for us and MAELSTROM has provided critical components to achieve this.

We are using the efficient and flexible data loaders developed in MAELSTROM for a lot of our work. We are also using many of the same benchmarking techniques from MAELSTROM when analysing the performance of our processing chains. The datasets stored in CliMetLab and the teaching material developed at the two MAELSTROM boot camps have been important for introducing machine learning techniques to the rest of our development team.

As an example, we are currently applying the same setup as in AP1 to improve the accuracy of our medium-range wind forecasts on Yr. In this case, we want to downscale and bias-correct the IFS medium-range ensemble from ECMWF using the same 1 km analysis system as we have done for temperature in AP1. Results are encouraging, showing that the model is able to correct IFS wind speeds in areas we know have strong systematic errors (Fig. 3).

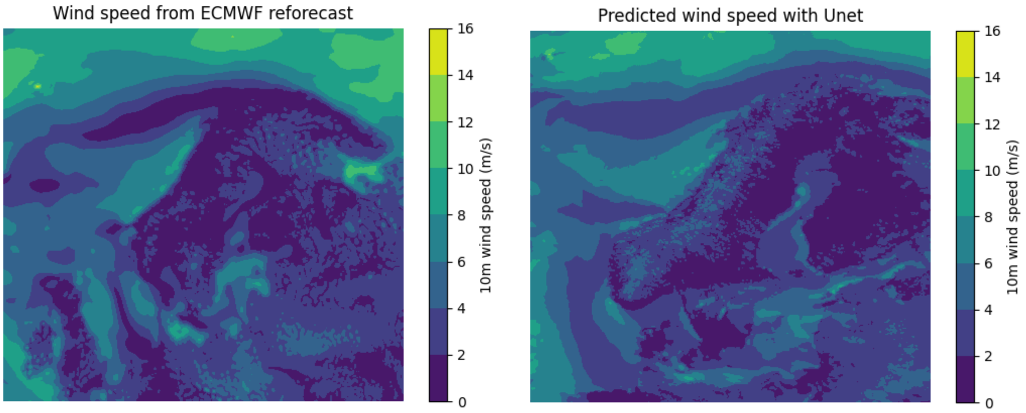

Wind speed from ECMWF reforecast
Predicted wind speed with Unet
10m wind speed (m/s)
0
2
4
6
8
10
12
14
16

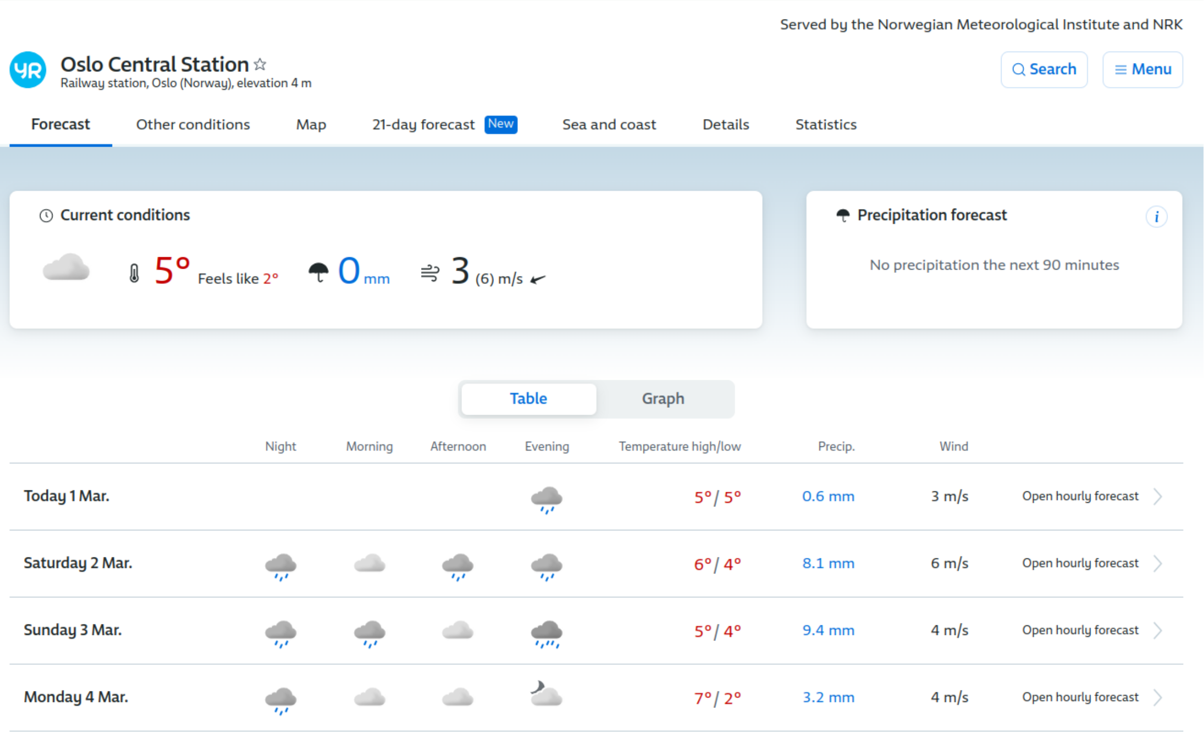
Fig. 3: Example of applying the AP1 U-Net to wind forecasts from the IFS ensemble, showing (a) 10m wind speed from the 0.1° resolution IFS used as input to the model, and (b) 10m wind speed output at 1 km resolution.

3.1.5 Impact

The temperature forecasts generated by the MAELSTROM project are available on https://www.yr.no for lookup locations in (mainland) Norway, Sweden, Denmark, Finland, the Baltics, and northern Germany (see Fig. 4 for an example). The improved forecasts were launched on Feb 27, 2024. In a typical week, more than 6 million unique Yr users will use the temperature forecasts developed by MAELSTROM. Short-range temperature is one of our most critical parameters for the general public, which greatly affects how users make decisions when planning their day.

Additionally, the data is made freely available on https://api.met.no (point forecasts) and https://thredds.met.no (gridded forecasts), which are used in a number of other apps and in automated systems by a wide range of downstream users making high-impact decisions within flood prediction, renewable power generation, transportation, and more. Our point API serves around 100 million requests per day, where a significant portion of the requests are for the domain where the MAELSTROM method has been implemented. The gridded files are requested over 100 million times per month and users download around 120TB of data every month.

It is important to say that the MAELSTROM project was essential for enabling MET Norway to develop and launch a deep-learning model for operational use. This includes the optimization of the data processing pipeline, the fine-tuning of the model architecture, and the hardware and software benchmarking techniques.



	Night	Morning	Afternoon	Evening	Temperature high/low	Precip.	Wind	
Today 1 Mar.				🌧	5° / 5°	0.6 mm	3 m/s	Open hourly forecast
Saturday 2 Mar.	🌧	☁	🌧	🌧	6° / 4°	8.1 mm	6 m/s	Open hourly forecast
Sunday 3 Mar.	🌧	🌧	☁	🌧	5° / 4°	9.4 mm	4 m/s	Open hourly forecast
Monday 4 Mar.	🌧	☁	☁	🌙☁	7° / 2°	3.2 mm	4 m/s	Open hourly forecast

Fig. 4: Screenshot showing temperature forecasts from MAELSTROM live on https://www.yr.no.

3.2 AP2: Incorporate social media data into the prediction framework

3.2.1 Background

In application two, we aim to harvest social media data to improve weather forecasts. Many posts on social media provide data that characterises the state of the weather at the current time and location of the user. This application aims to extract relevant information and use it as additional observations to be integrated into the data assimilation process. As an initial test case, we attempt to deduct the presence of rain from Tweets. For this, we accumulate 1.4 Mio Tweets that contain keywords related to precipitation (e.g., “rain”, “downpour”, “sun”, …) and labelled them as “raining” or “not raining” based on ERA5-land data (see D1.4 for details). We constrain Tweets to the English language that were sent from the UK.

3.2.2 Further developments since D1.4

In D1.4, we identified data quality as the main limiting factor for performance. First, we noticed that the precipitation dataset ERA5-land used for labelling our Tweets showed clear deviations from our dataset derived from measurements at weather stations. To tackle this, we built a small holdout dataset of Tweets that are near weather stations. We use this holdout dataset as the final performance measurement of our model as we deem the labels most accurate. In addition, we noticed that a significant fraction of Tweets do not currently provide sufficient information (even for humans) to correctly state if it was “raining” or “not raining” at the place and time of the sender of the Tweet. This is due to our current keyword based filtering, which is deemed insufficient for these constraints. To remedy this issue, we suggested an additional classification model that labels Tweets as “relevant” or “not relevant” regarding our task. If deemed “not relevant”, the Tweet contains insufficient information for the rain classifier to make its prediction and it is discarded from the dataset.

Since D1.4, we implemented an initial relevance classifier. We used a text generating LLM Falcon-40b to generate a training dataset for the relevance classifier by prompting it to label individual Tweets as “relevant” or “not relevant”. By experimenting with training sizes, we determined that we require at least 20k Tweets to train our rain classifier. We therefore generated labels for 25k (5k each for testing and validation) Tweets with Falcon to train our relevance classifier. We realised that a significant fraction of generated text did not match the desired format and/or content, which we discarded. The prompt responses seemed promising. However, during evaluation of the relevance classifier, we found performance insufficient. Further comparison with a hand-labelled dataset of 100 Tweets determined that our best performing prompt was only able to label 30 Tweets correctly (50 Tweets had to be discarded due to incorrect formatting).

We aim to test additional models and prompts. Additionally, constraining the format in which the LLM generates text should allow us to retain more responses.

3.2.3 Requirements for potential implementation

The final model is expected to estimate the presence of rain at the location and time of social media users. This information is interpreted similarly to a conventional observation by a weather station or weather satellite. Therefore, their integration into a potential forecasting system would require similar steps. Data assimilation describes the process of creating the initial conditions used for numerical weather predictions based on weather measurements. However, as precipitation is conventionally measured and predicted as floating point values, a new data category would have to be devised including a procedure to align our binary prediction with the floating point values.

In addition, Tweets correspond to a new data source compared to conventional measurements. Globally on the order of 500 million tweets are sent daily. If all Tweets should be considered, a data stream should be set up that selects “relevant” Tweets. Initially, keyword based filtering could remove the vast majority of Tweets. Further filtering is provided by the relevance classifier. For the UK, we expect on the order of 1000 Tweets would pass the keyword filtering applied by us, which is a manageable dataset for running inference both with the relevance and subsequently the rain classifier. Afterwards they could be included in the data assimilation process or in post-processing systems.

This small amount of observations is not expected to have a significant impact on the overall performance of the prediction of precipitation. However, in areas with higher concentration of Twitter users like urban areas, performance gains are expected. As many consumers of weather forecasts are expected in these areas, one may expect performance gains on neighbourhood or even street level with the proposed method.

To include more Tweets in the analysis, keyword-based filtering would have to be abandoned as an initial step. This would require the challenging setup of smarter pre-filtering techniques as otherwise model inference may become a significant bottleneck.

In the future, additional data sources could be considered, including other data types like images and/or videos. For example, images provided by Twitter or Instagram could be included in the analysis or even videos from TikTok. This would require the development of new model architectures.

3.2.4 MAELSTROM workflow tools

The newly devised MAELSTROM workflow tool Mantik relies heavily on MLflow. MLflow provides excellent features for tracking, comparing and managing results from ML models. This feature aided the development of this application considerably. As model building and training is a major workflow of our company, we are very interested in new tools that help our teams. We are therefore planning the implementation in our workflow at 4cast.

3.3 AP3: Build neural network emulators to speed-up weather forecast models and data assimilation

3.3.1 Background

Application three concerns training neural networks to emulate computationally expensive components of weather forecasting models, with the purpose of creating cheaper but minimally less accurate components. In prior work, this has been demonstrated as a proof-of-concept for smaller components of the weather forecasting chain. Here we target a particularly complex physical process, radiative transfer. This process captures longwave and shortwave heating and cooling, involving interactions with clouds, aerosols and the land surface.

Two years of global training data was generated, covering diurnal and seasonal cycles.

3.3.2 Further developments since D1.4

In previous deliverables, outstanding accuracy was achieved with solutions for both longwave and shortwave processes. Both required custom architecture designs, heavily motivated by the physical process and numerical solver involved in the radiative transfer scheme. This accuracy was measured offline, i.e. decoupled from the IFS, the weather forecasting model run by ECMWF. The true test is coupling the models into the IFS, replacing the numerical solver for radiative transfer, and then assessing the accuracy of the weather forecasts. Work has focussed on effectively coupling complex neural networks into a Fortran-based weather forecasting and then assessing the accuracy. This will now be expanded upon.

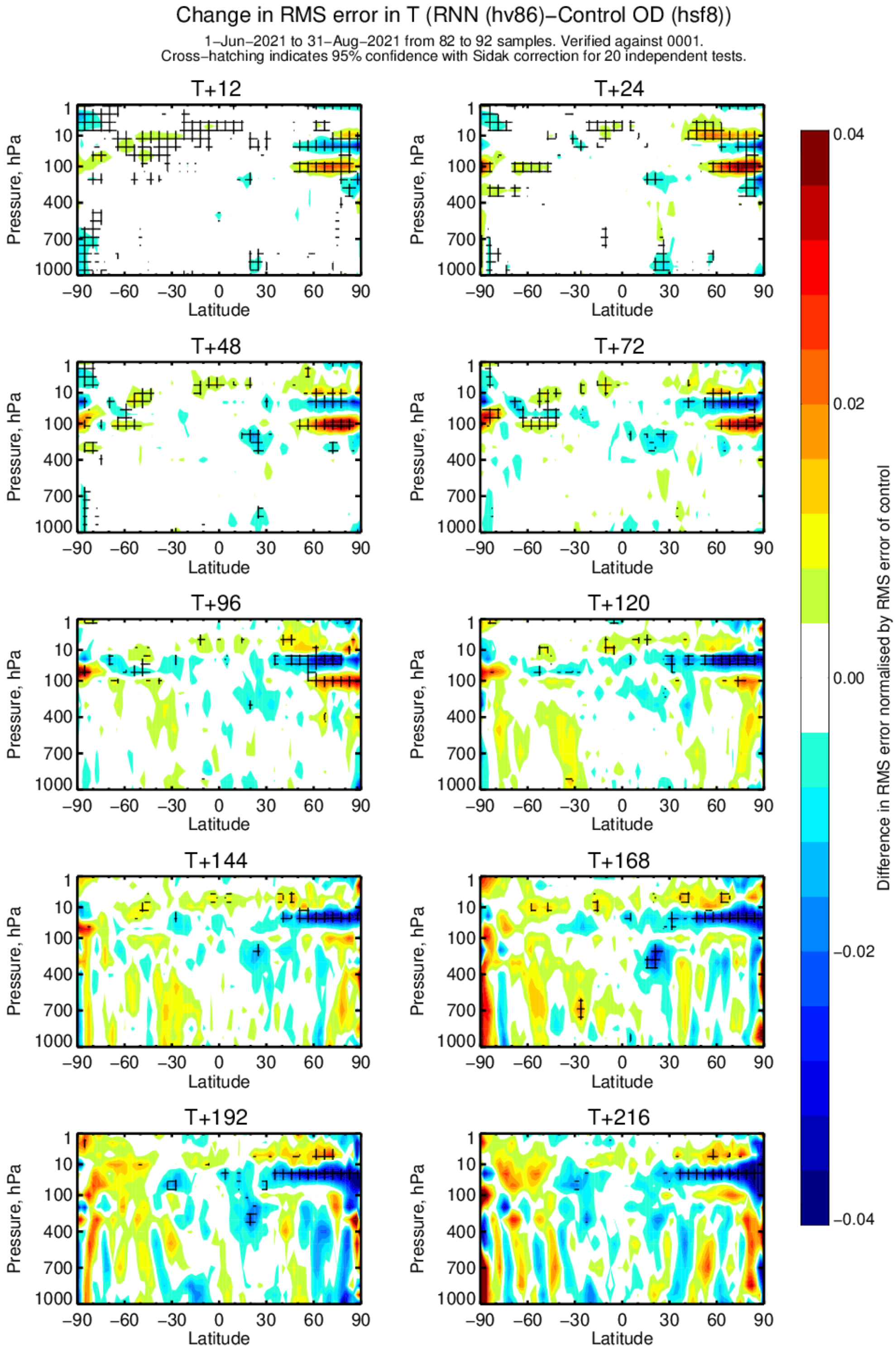
3.3.3 Coupling neural networks into the IFS

Coupling was achieved using the software package Infero, which ECMWF has developed coupling machine learning into Fortran. The software is open source, and available at https://infero.readthedocs.io/en/latest/index.html. The solution provides a standardised interface to multiple inference backends, including ONNX-runtime, Tensorflow LITE, Tensorflow-C and TensorRT. The radiative transfer emulators were converted from Keras to ONNX-runtime objects. Next, a small module was developed in the IFS to load the neural networks into memory, package the inputs and then pass them into the neural network.

3.3.4 Testing emulators within the IFS

To test the radiative transfer emulators, a series of forecast experiments were run, where the physics-based radiative transfer model was switched off, and the two neural networks for longwave and shortwave processes were used in place. Forecast experiments were run for June-July-August 2021 and December-January-February 2021/22, providing data for summer and winter seasons. These are independent years from the year 2020, which was used for training. The forecasts were run using the TCo399 configuration of the IFS, which has a resolution of approximately 0.25 degrees and is the standard resolution for testing forecast developments. In Fig. 5, results from these forecasts are plotted. Overall they show high quality emulators have been trained, and even in online testing they show almost no change in the forecast accuracy. The statistically significant changes are a mix of positive and negative, meaning overall a neutral change in forecast quality. From an accuracy perspective, these emulators would be accurate enough to consider use in an upcoming operational cycle.

Accuracy is a necessary condition for the emulator to be considered for operational use, but not sufficient. For that purpose the emulator should be significantly faster than the original component, ideally significantly faster. With the Infero library, inference within the IFS can be carried out on either CPU or GPUs. Initial exploration of relative speed, when the radiation emulators were deployed on GPU showed significant acceleration over the conventional physics-based radiation scheme, deployed on CPU. In certain configurations, the emulator was approximately five to ten times faster. In parallel, and outside of the MAELSTROM project, significant work was invested to improve the computational performance of the physics-based solver. This had a dramatic impact, improving the computational performance of the ecRad radiation scheme by up to twelve times (Ukkonen and Hogan 2024). The net effect is that for this version of the radiative transfer emulator, no significant acceleration is gained. This is an important result for the field, as there is a presumption that ML models will accelerate calculations. This discovery suggests that if the physics model does not have a significant enough computational burden, then acceleration will be challenging. However, there are still ways to exploit these results and still improve weather forecasts.


Change in RMS error in T (RNN (hv86)–Control OD (hsf8))
1–Jun–2021 to 31–Aug–2021 from 82 to 92 samples. Verified against 0001.
Cross–hatching indicates 95% confidence with Sidak correction for 20 independent tests.
T+12
T+24
T+48
T+72
T+96
T+120
T+144
T+168
T+192
T+216
Pressure, hPa
1
10
100
400
700
1000
Latitude
–90
–60
–30
0
30
60
90
0.04
0.02
0.00
–0.02
–0.04
Difference in RMS error normalised by RMS error of control

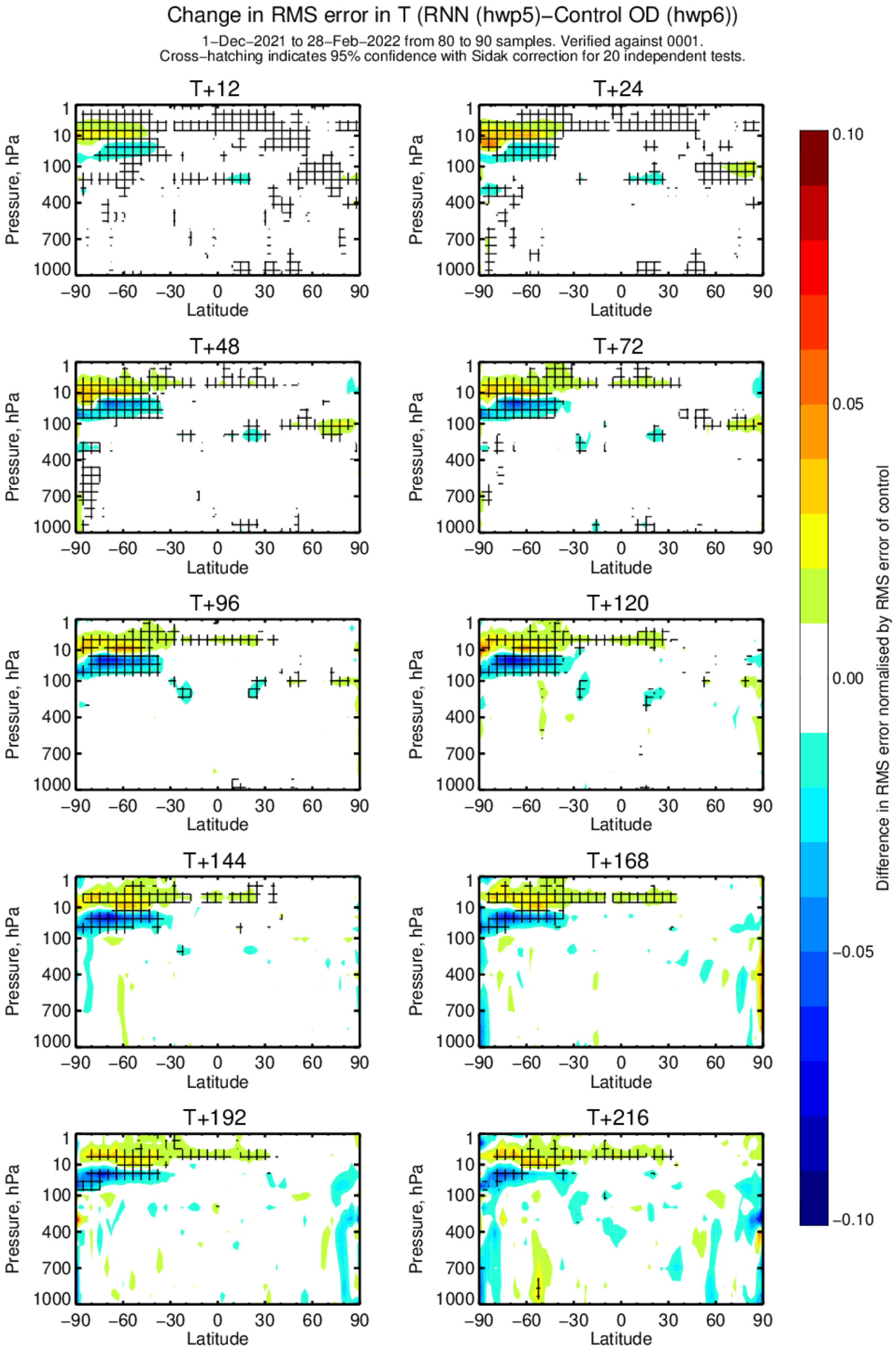


Fig. 5: Change in forecast accuracy for atmospheric temperature when using neural network emulators for radiative transfer, with hatchings denoting statistical significance. Red shows forecast degradation when using the emulator, blue improvement. Almost no statistically significant change is seen, with a mix of positive and negative impact over the summer pole.

3.3.5 Further exploitation of results at ECMWF

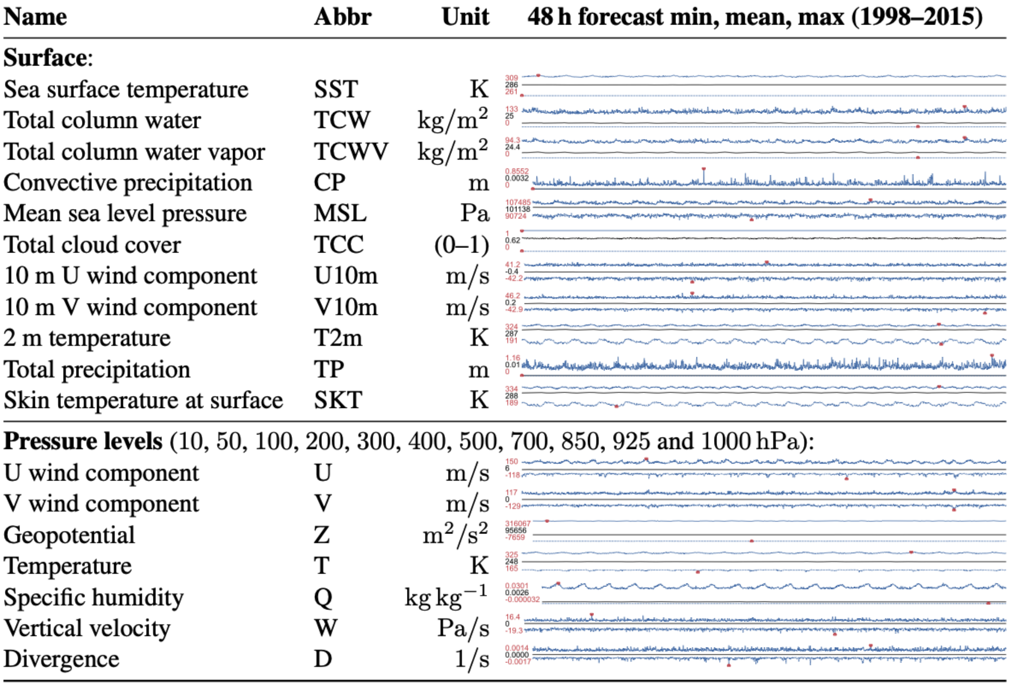
Despite the lack of acceleration for the current radiation emulator, there remain exciting prospects for this approach to improve forecast quality.

Firstly, there exist more complex radiative transfer solvers, which incorporate more physical processes and fewer approximations, such as the SPARTACUS solver (Hogan et al. 2016). This has a significantly higher cost (between 10 and 100 times more expensive), currently too expensive for use in an operational weather forecasting model. Within the MAELSTROM project, training, testing and validation datasets from the SPARTACUS solver have been generated, but not yet used for model development. Beyond this project, these datasets will be used to train and test emulators of this even more expensive solver, with the aim to make it affordable for operational use. Work in MAELSTROM highlights that accurate and stable forecasts with ML emulators of radiative transfer is possible, and provides a recipe for success.

Secondly, as discussed in D1.5, in addition to the acceleration angle, ML emulators of physical parametrisation schemes could have a role for the automatic generation of derivative code for use in data assimilation applications. Results in deliverable D1.5 show exciting promise for this use, with the initial tests passed for the more complex ML solutions. Further work will be carried out after MAELSTROM to test in more detail the validity of this approach, including testing the tangent linear and adjoint models within the full data assimilation algorithm. Should these tests prove successful, emulators would likely be developed for many areas of parameterized physics.

3.4 AP4: Improve ensemble predictions in forecast post-processing

Application four aims to make weather forecasts better by using deep neural models to refine the predictions from ensemble weather models. In the first step, we introduced a new dataset called ENS-10 (Ashkboos et al, 2022), which includes ten different forecasts covering 20 years from 1998 to 2017. These forecasts are created by weather simulations to capture the Earth's unpredictable behaviour and provide a reference for climate states at a certain date. ENS-10 gives us a detailed view of the atmosphere with various key weather variables at different pressure levels and the surface, and it's available for forecast lead times of 0, 24, and 48 hours with two forecasts started per week. Our main goal is to improve 48-hour forecasts by correcting them using the ENS-10 dataset.



Name	Abbr	Unit	48 h forecast min, mean, max (1998–2015)
Surface:			
Sea surface temperature	SST	K	309 286 261
Total column water	TCW	kg/m^2	133 25 0
Total column water vapor	TCWV	kg/m^2	94.3 24.4 0
Convective precipitation	CP	m	0.8552 0.0032 0
Mean sea level pressure	MSL	Pa	107485 101138 90724
Total cloud cover	TCC	(0–1)	1 0.62 0
10 m U wind component	U10m	m/s	41.2 -0.4 -42.2
10 m V wind component	V10m	m/s	46.2 0.2 -42.9
2 m temperature	T2m	K	324 287 191
Total precipitation	TP	m	1.16 0.01 0
Skin temperature at surface	SKT	K	334 288 189
Pressure levels (10, 50, 100, 200, 300, 400, 500, 700, 850, 925 and 1000 hPa):			
U wind component	U	m/s	150 6 -118
V wind component	V	m/s	117 0 -129
Geopotential	Z	m^2/s^2	316067 95656 -7659
Temperature	T	K	325 248 165
Specific humidity	Q	kg kg^{-1}	0.0301 0.0026 -0.000032
Vertical velocity	W	Pa/s	16.4 0 -19.3
Divergence	D	1/s	0.0014 0.0000 -0.0017

Fig. 6: Parameters of ENS-10. The dataset contains each parameter for 10 different ensemble members on resolution 0.5° latitude/longitude grids at 0, 24, and 48 hour forecast lead times.

In the next step, we assessed the prediction accuracy of different deep-learning models using ENS-10. To this end, we evaluate using “Simple” learned statistical methods, such as EMOS, which are also broadly representative of what is presently used by production forecast systems for post-processing; MLPs; simple convolutional neural networks; U-Nets; and transformers on ENS-10 dataset. We evaluate the above models using CRPS and EECRPS (which is essentially the CRPS metric, weighted by extreme forecast index) metrics.

3.4.1 Production Usage of AP4

To make our application ready to be used in production, we open-sourced all our code and data. More precisely, ENS-10 is available under the Creative Commons Attribution 4.0 International (CC BY 4.0) licence. The scripts to get the dataset, run (and edit) the models, train, benchmark, and evaluate are available in the ENS-10 GitHub repository[1]. In addition, we also provide a CliMetLab plugin for downloading the dataset for a fixed time and variable[2].

Our application could be used at the production level in two different ways: First, one can use our publicly available dataset to develop new methods for post-processing weather forecasts. In addition to the dataset usage, one can use our models for post-processing weather forecasts and compare them against new (or other) models as a benchmarking tool.

Currently, the main limitation of using our application at the production level is the limits of the dataset. Our dataset contains only 30 years with 0.5-degree resolution. To apply the tool directly to

[1] https://github.com/spcl/ens10/tree/main

[2] https://github.com/spcl/climetlab-maelstrom-ens10

operational global ensemble predictions at ECMWF at the native resolution of 9 km, one should readjust the tools to the higher resolution grids.

3.4.2 Further exploitation of results at ETHZ

ETHZ is hosting the ENS-10 dataset, allowing both internal and external users to integrate it into their research or production, and we plan to continue hosting the dataset long after the MAELSTROM project completes. It has proven to be a valuable dataset in research at ETHZ on compression for weather and climate data, diffusion models for climate data assimilation, and other ongoing work. It has been used by a number of many external collaborators as well as ETHZ students for thesis projects and project work. Since its publication in 2021, the paper describing the ENS-10 dataset has been cited 15 times.

We will continue using the ENS-10 dataset in our research and education i.e., for student thesis projects. At the same time we apply the experience gained from MAELSTROM in follow-up projects. In the future we will attempt to generate and work with bigger datasets (longer time-period and higher resolution).

3.5 AP5: Improve local weather predictions in forecast post-processing

3.5.1 Background

In application five, we developed deep neural networks for accurate statistical downscaling of meteorological fields. Downscaling of 2m temperature constitutes the central task of this application, since the near-surface temperature has a high societal and economic relevance (e.g. for agriculture) and since it exhibits a high degree of spatial variability in the presence of complex terrain.

Originally, the application was aimed towards forecast post-processing and thus, the first benchmark dataset was based on short-range forecasts of the IFS, the operational numerical weather prediction (NWP) model at ECMWF. However, the accuracy of data-driven deep neural networks crucially depends on the availability of consistent and large-scale training datasets. For NWP data, the fulfilment of this requirement is often complicated by model physics and resolution updates in the operational forecast chain, irregularly applied over the years. For instance, the operational IFS model underwent a major horizontal resolution update in March 2016 and a revision of the moist physics parameterization in October 2021 in addition to several smaller updates in between.

By contrast, consistent, multi-year datasets can be created with the help of reanalysis datasets which are generated with a fixed model version. The ERA5 and the COSMO REA6 datasets constitute well-established and quality-controlled reanalysis datasets that are frequently used to monitor climate change or to assess the potential of renewable energy production (Jourdier, 2020;). Both reanalysis datasets have therefore been chosen as the base of the advanced second dataset for 2m temperature downscaling. The added value due to the finer grid of the COSMO REA6 data ($\Delta x_{CREA6} \simeq 6$ km) compared to the ERA5 ($\Delta x_{ERA4} \simeq 30$ km) have been verified in various studies, e.g for temperature over complex terrain (Scherrer, 2020).

Thanks to the comprehensive temporal coverage of the ERA5-reanalysis (between 1940 and near present), the downscaling model for 2m temperature of Application 5 can in principle be used for a temporal extension of the COSMO REA6-data which is limited to the period between 1995 and August 2019. Progress on the developed model solution since the recent deliverables is provided below. This is complemented by a detailed description of the exploitation plans at JSC, including future work on the large-scale Foundation Model AtmoRep probed in D1.4.

3.5.2 Developments since D1.3 and D1.4

Since the reports in the recent deliverables of work package 1, the dataset and the Wasserstein Generative Adversarial Network (WGAN) solution for the 2m temperature downscaling task have been improved. Dataset enhancements include a change from pressure-based to modellevel-based temperature predictors, as well as augmenting the dataset by incorporating additional data from 1995 to 2006 as available with the COSMO REA6 dataset. The WGAN has been improved by changing the upsampling layer and switching from ReLu to the Swish activation function in all convolutional layers of the U-Net generator and critic model component. In summary, the updates led to a nearly 20% reduction in the RMSE.

Using modellevel-based temperature predictors. As documented in D1.3, larger errors have been noted for the Alpine region, especially during the summer and spring months around noon. This deficiency could be attributed to the pressure-based predictors (the 850 hPa- and 925hPa-temperature), which provided insufficient information to reconstruct the temperature on elevated grid points above the 850 hPa-level ($\gtrsim$ 1500m). While the tested date embeddings only yielded small improvements in D1.3, a significant reduction in the RMSE was achieved by incorporating the temperature from the model-levels 137, 135, 131, 127, 122 and 115 of the ERA5 dataset to the set of predictors. Specifically, the (domain-)averaged RMSE for 11 UTC during the summer months (June, July, August) decreased from 1.64 ± 0.47 K to 1.38 ± 0.23 K, mainly due to reduced errors over highly elevated terrain. The overall average on the test dataset for 2018 improved to 1.07 ± 0.26 K (compared to $RMSE = 1.15 \pm 0.31$K). Albeit a minor reduction of the gradient amplitude ratio r_{WGAN} can be analysed in comparison to the date embedding approach ($r_{VT} = 0.950$ vs. $r_{VT}^{embed} = 0.970$), this indicates that physical, informative predictors are most efficient for decreasing daytime- and season-dependent degradations in accuracy.

Tuning the WGAN architecture. Based on feedback discussions with the developers of AP1 and AP3, a systematic tuning of the WGAN architecture was conducted where changes to the pooling layer of (average vs. maximum pooling) and the upsampling layer (bilinear upsampling vs. upsampling via the subpixel layer (Shi et al., 2016)) of the U-Net generator have been tested. Furthermore, the activation functions LeakyReLu and Swish have been probed since they may improve the performance and the generalisation capacity of deep neural networks compared to the ReLu activation function that has been initially used (Maas et al., 2013; Ramachandran et al., 2017). Experiments show that the usage of maximum pooling and the subpixel layer for down- and upsampling in the U-Net generator as well as choosing the Swish activation can further boost the downscaling accuracy in terms of the RMSE by about 7%. The average RMSE on the test dataset gets reduced to 1.00 K, while the spatial variability remains largely unchanged. Noticeably, average pooling improves the stand-alone U-Net model (not shown), whereas no benefits in terms of RMSE, but a slight degradation in reproducing the spatial variability has been noted for the WGAN. The results of the tuning experiments are displayed in Fig. 7.

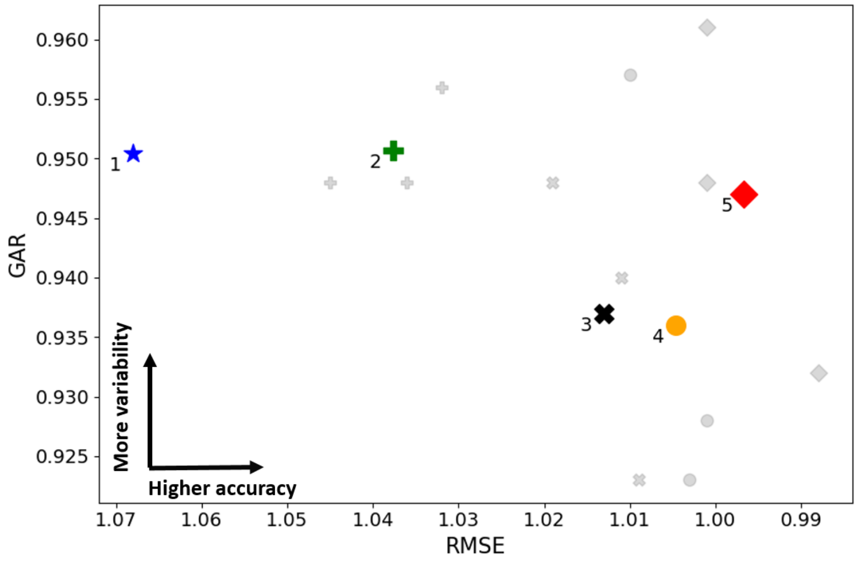


Fig. 7: Averaged gradient amplitude ratio (r_{GAN}) plotted against RMSE for the WGAN tuning experiments. The blue star (1) represents the untuned WGAN attained after training with the dataset comprising modellevel-based temperature predictors. The average results of the WGAN experiments using a subpixel layer for upsampling and average pooling together with the ReLu activation function are shown by the green cross (2). The black cross (3) and the orange circle (4) are similar, but use the LeakyReLu and Swish activation function, respectively. The final tuned WGAN configuration with max-pooling and Swish is denoted by the red square (5). The small grey markers represent the individual experiments for the probed configurations.

Completion of the downscaling dataset. The Tier-2 dataset presented in D1.3 just comprised data between 2006 and 2018. Recently, the dataset has been completed and now involves data back to 1995, the first year for which the COSMO REA6-analysis is available. The corresponding CliMetLab plugin (https://pypi.org/project/climetlab-maelstrom-downscaling/, more details on CliMetLab are provided below) has also been updated accordingly.

Increasing the amount of samples to 192,576 yields another boost in terms of the RMSE. Evaluated on the test dataset, another decrease in the RMSE by about 4% ($RMSE_{WGAN} = 0.962 \pm 0.248$ K) is attained. The RMSE reduces now well below 0.9 K for the evening and night hours, whereas a maximum with an RMSE around 1.2 K persists around noon (Fig. 8a). This maximum is still most pronounced in the spring and summer season, when the spatial variability of the 2m temperature over complex terrain is largest (Fig. 9b) in a convective planetary boundary layer (PBL). On average, the spatial variability is slightly underestimated with an average gradient amplitude ratio $r_{VT} = 0.953 \pm 0.067$ (Fig. 8b). This result is confirmed by analysing the power spectra of the 2m temperature field (not shown). In the summer months, this underestimation is strongest, whereas an overestimation of the spatial variability is diagnosed for the spring months (not shown). This underpins the difficulties in accurately reconstructing the local 2m temperature field when the PBL is well mixed. The exact capturing of the near-surface lapse rate driving the spatial variability over complex terrain is challenging since it depends on the land use and the soil moisture that ultimately control the surface fluxes on local scales. This information is probably not reliably encoded with the surface flux predictors from the ERA5 dataset.

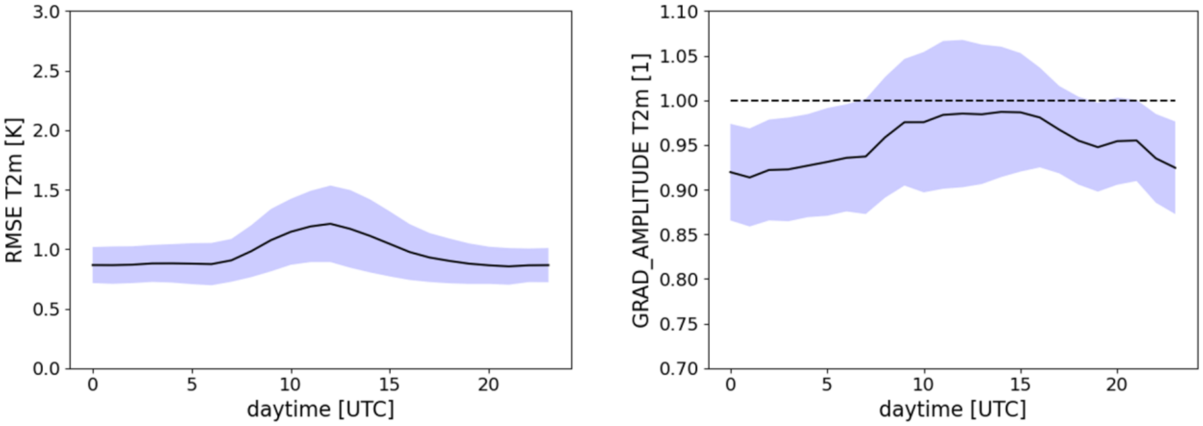


Fig. 8: (a) Domain-averaged diurnal cycle of the RMSE and (b) the gradient amplitude ratio r_{GAN} evaluated over the whole test data from 2018. The dashed line in (b) denotes the optimal value in which the average amplitude of the T2m-gradient is the same in the downscaling product as in the ground truth data from COSMO REA6.

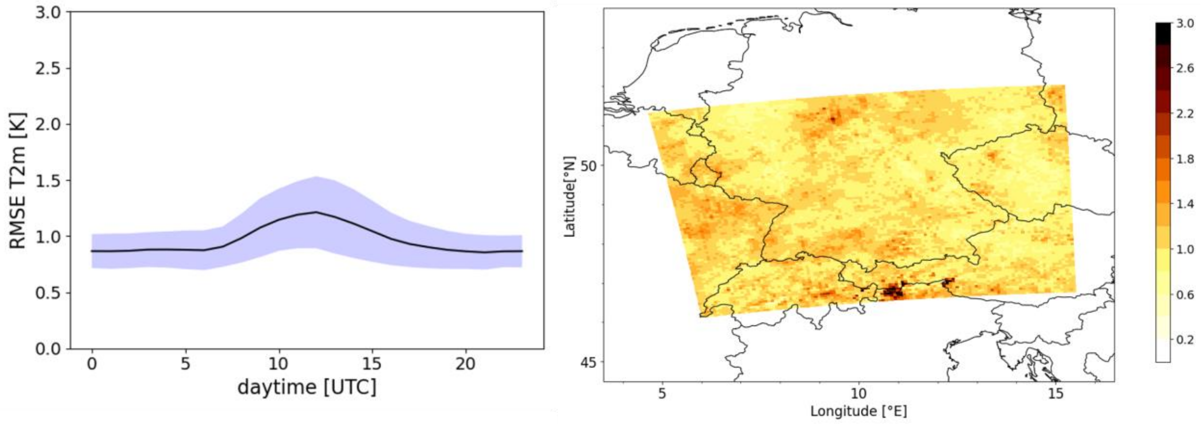


Fig. 9: (a) Domain-averaged diurnal cycle of the RMSE for the summer months June, July and August (JJA). (b) Spatial distribution of the RMSE evaluated at 12 UTC for the summer months JJA.

3.5.3 Updates on AtmoRep

Since the last deliverable, downscaling with AtmoRep has not undergone significant improvements. Instead, a preprint of the AtmoRep-paper in which downscaling is described as one of the downstream-applications has been published on arXiV (Lessig et al, 2023). Furthermore, the AtmoRep model source code and relevant analysis code have been published via two github-repositories[3]. The github-repository for the AtmoRep source code also provides guidelines to download the training data as well as different pretrained AtmoRep models of which the Multiformer configuration `multi3-uv` (trained on temperature and the horizontal wind vector components) corresponds to the AtmoRep variant that has been used for the downscaling application.

[3] https://github.com/clessig/atmorep/ and https://github.com/iluise/atmorep_analysis

3.5.4 Further exploitations of results at FZJ

3.5.4.1 Towards a benchmark dataset for statistical downscaling

During the course of the MAELSTROM project, the field of downscaling with deep neural networks has seen rapid progress. While Generative Adversarial Networks such as the WGAN developed in this application have become standard and well-established approaches (e.g. Price and Rasp, 2022, Hess et al. 2023, Vosper et al., 2023), diffusion models have recently become state-of-the-art for a wide range of downscaling tasks (Hatanka et al., 2023, Mardani et al., 2024). Despite the encouraging results presented in several studies, a comparison between the growing number of approaches is impeded due to the vast amount of different datasets that are applied for downscaling and due to varying metrics used for evaluation. A way to ease intercomparison between different ML solutions are benchmark dataset, which have steered progress over the last decade in several AI domains, such as large language modelling and computer vision. While benchmark datasets like WeatherBench (Rasp et al., 2023) and ClimateBench (Watson-Parris et al., 2022) have been released for medium-range forecasting and climate applications, the downscaling field still lacks a well-defined framework and standard dataset.

The team of AP5 has therefore developed the idea of providing such a benchmark exploiting the work pursued in the scope of MAELSTROM. Like Application 5, the benchmark dataset will be based on the ERA5- and COSMO REA6-reanalysis to provide ready-to-use databases for downscaling the 2m temperature, the 100m-wind and the global horizontal irradiance. The data preprocessing chain developed in MAELSTROM is used to create the datasets for the three downscaling tasks, while the evaluation framework will also use components established during the project (e.g. the analysis in terms of the gradient amplitude ratio $r_{\nabla T}$). Furthermore, the U-Net and the WGAN architecture of AP5 will be integrated as baseline model solutions, complemented by the DeepRU U-Net variant suggested in Hoehlein et al. (2020) and the transformer-based SwinIR-architecture (Liang et al. 2021). The Standardized Anomaly Model Output Statistics (SAMOS) by Dabernig et al., 2017 will serve as an additional competitor model to assess the added value of ML-based solutions against classical statistical approaches.

While the preparation for the publication of the benchmark dataset including an accompanying paper are on-going and planned for late spring, some preliminary results are already available as shown in Fig. 10.

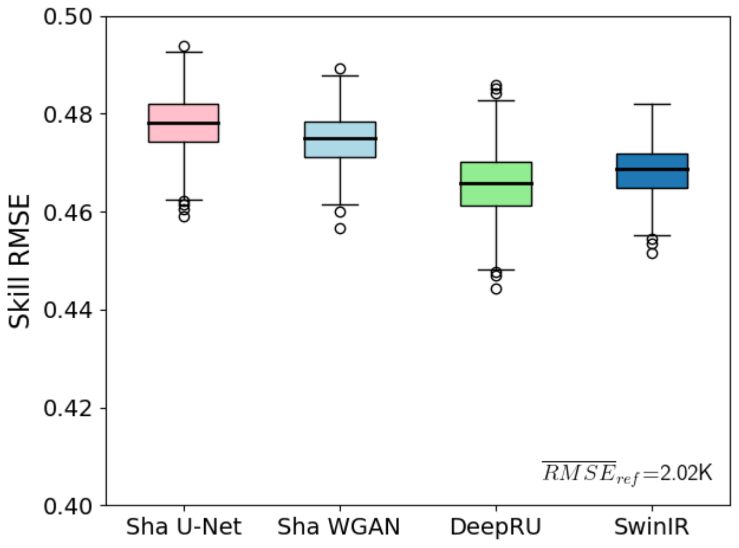


Fig. 10: RMSE skill score analysis for different baseline ML downscaling models. The Sha U-Net (Sha et al., 2020) is used as the generator model in the WGAN and has been developed and tuned in this application. The DeepRU and SwinIR architectures are adapted from Hoehlien et al., 2020 and Liang et al., 2021, respectively. Bilinear upsampling serves as the reference 'downscaling' method for calculating the RMSE skill score.

It is noteworthy that this work will be continued in a collaboration between JSC and Geosphere Austria - an initiative that was forged during the recent MAELSTROM BootCamp in Reading (cf. D4-8). A further connection was established with the Mila Quebec AI Institute represented by Paula Harder who has been working successfully on physical constraints for downscaling (Harder et al., 2023). Thus, the MAELSTROM's dissemination activities have substantially contributed to push international collaborations to ultimately advance the field of statistical downscaling with deep neural networks.

3.5.4.2 Downscaling of air pollutants in the DestinE use case DE_370c

The developed WGAN model of AP5 is also exploited in the scope of the Destination Earth use case DE_370c. While Destination Earth aims to develop highly-accurate digital twins of the Earth System, the use case application focuses on air quality forecasting based on the high-quality atmospheric simulations of the extreme digital twin. The central objective of the use case is to develop a user interface for on-demand air quality forecasts by combining the numerical chemical transport model EURAD-IM with machine-learning applications. The German environment agency UBA and the North Rhine-Westphalia Office of Nature, Environment and Consumer Protection (LANUV) constitute the core user of the interface.

To provide on-demand air quality forecasts at kilometre-scale with minimised latency time, different WGAN models are trained to downscale air pollutant forecasts for ozone, nitrogen dioxide, PM2.5 and PM10 from the EURAD-IM model. The EURAD-IM model has been run operationally since 2012 with a high resolution nested domain ($\Delta x = 1$ km) over North-Rhine Westphalia, enabling the creation of a multi-year training dataset for ML downscaling models. The downscaling product should thereby support the core user's decision to run computationally expensive EURAD-IM simulations at 1 km resolution if required due to high-impact local air pollution events (e.g. smog). This is possible since the inference with trained deep neural networks requires little computational resources, providing the results in less than one minute.

While the downscaling models are still under development, prototype solutions already exist. Downscaling air quality data however imposes new challenges, especially for quantities that are strongly driven by local emissions such as nitrogen oxide (NOx) as shown in Fig. 11. The experience gained in the MAELSTROM project is thus of great relevance to increasing the accuracy of the developed downscaling models and to leverage the available training data from historical EURAD-IM simulations.

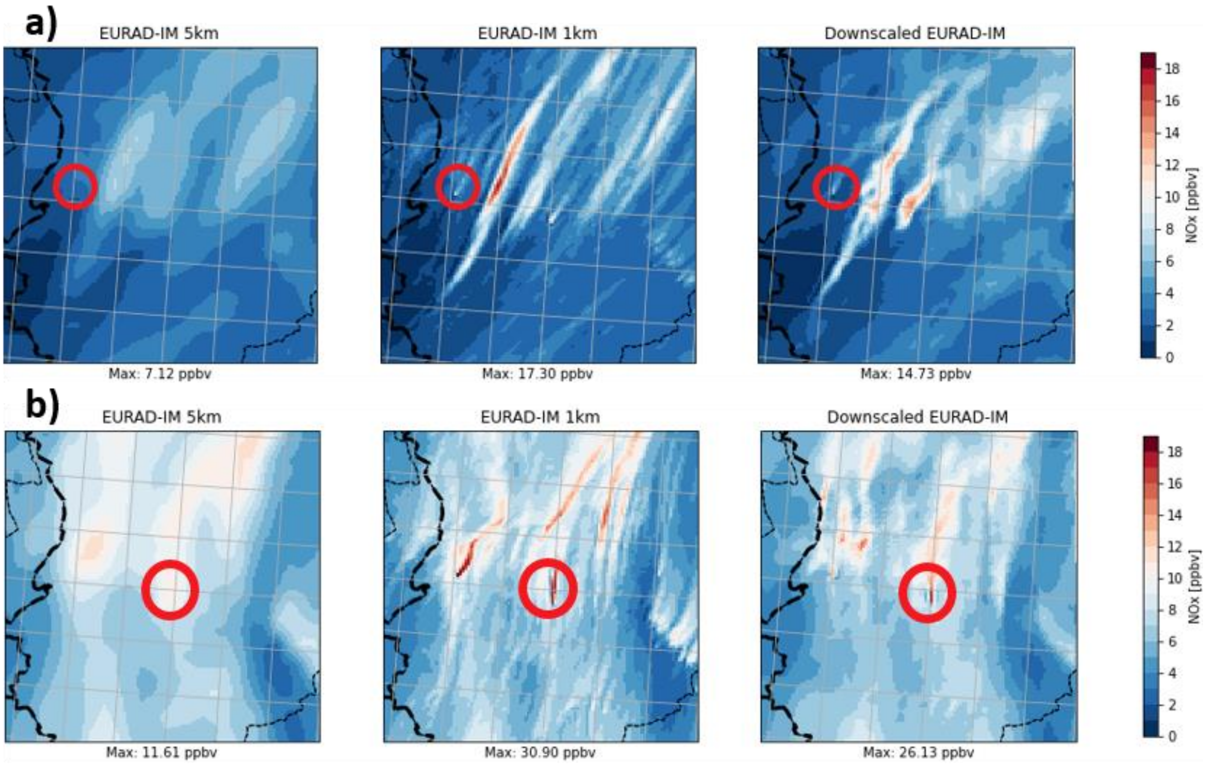


Fig. 11: Preliminary results of downscaling NOx EURAD-IM forecasts with the WGAN architecture developed in AP5 for (a) 2018-01-01 11 UTC and (b) 2018-02-11 17 UTC. The coarse-grained input on a 5km-grid is shown in the left column, while the targeted 1km EURAD-IM data is shown in the centre. The downscaled NOx-field is shown on the right.

3.5.4.3 Future developments of AtmoRep

The seminal paper of Lessig et al. (2023) has shown that Foundation Models for atmospheric dynamics have the potential to push the frontiers of large-scale neural networks in weather and climate applications. While AtmoRep has been solely trained on ERA5 reanalysis data yet, future work is dedicated to incorporate various Earth system datasets, including satellite and synoptic observations, along with numerical model outputs. The aim is to enable very large scale representation learning of the Earth system’s state, accommodating a wide range of spatio-temporal scales and leveraging incomplete and partially inconsistent data sources. The refined Foundation Model will be tested across various applications, of which several require high-resolution data such as the prediction of local precipitation and extreme wind events. Hence, statistical downscaling to kilometre-scale and beyond will remain central to the development of AtmoRep and its successor, enabling the emergence of state-of-the-art ML solutions, even when confronted with limited availability of large-scale, consistent training data for specific tasks as argued above.

3.6 AP6: Provide bespoke weather forecasts to support energy production in Europe

In application six, we developed ML models for power production forecasts of renewable energy (wind and solar). In MAELSTROM, we have investigated whether recurring large-scale weather regimes (LSWRs) can yield information about the power production and the uncertainties associated with their forecasting, and whether the information can be fed back into the ML models to improve the forecast uncertainty with a prior knowledge of occurring weather regimes.

We've followed conventional linear dimensionality reduction methods as well as nonlinear Deep Learning (DL) algorithms (Caron et al. 2020) to first see whether we can build a model that finds a set of recurring LSWRs that correlate with the power production and/or forecast uncertainty. This model could then be used as a classifier to predict occurring LSWRs based on the output of operational weather models such as the IFS. These predicted LSWRs could potentially be incorporated into our current ML models to improve the forecast uncertainty due to the large-scale weather information contained in the new input feature produced by the classification model.

3.6.1 Incorporating Large-Scale Weather Regimes into the Forecasting Models

There still is ongoing research whether using the recurring LSWRs as input features to our models may increase the model performances, and how the improvements affect the performance of different model architectures. We also want to further explore to what extent the correlation between LSWRs and forecast uncertainty may help estimating forecast uncertainty in advance, and whether there is a potential market for such forecasting products.

4cast provides operational forecasts for renewable power generation to a range of customers. If the classification model turns out to increase model performance, we will integrate the tools into the forecasting system, i.e. the current training and forecasting pipelines of our ML models that deliver forecasts for our customers.

3.6.2 Usage of the Workflow Tools

For the entire process pipeline, we've made extensive use of Mantik. Our research application has multiple entry points, each with a different scope (data preprocessing, conventional ML algorithm, deep learning model training), and Mantik allows us to run each of these entry points on any of the supercomputers at the Juelich Supercomputing Centre and to supervise the results (i.e. input parameters, evolution of loss, metrics, and plots) in real time. To achieve this, the application makes use of MLflow, which is embedded in the Mantik platform, for tracking the training process and exploring results.

For benchmarking, we've also integrated the Deep500 framework into the application to track the performance. Deep500 works in conjunction with MLflow and allows us to track the measured performance metrics directly to the Mantik platform.

3.7 AP8: Operational running of machine learning weather forecasting models

3.7.1 Background

During the course of MAELSTROM, a new prospect emerged for operational weather forecasting. Whilst the topic of completely data-driven machine learning forecasting systems had been proposed in 2018 (Dueben & Bauer 2018), the results over the following few years did not suggest this would be a viable approach. The models are trained from the global reanalysis datasets ERA5 and use pure machine learning approaches to build models of the global atmosphere that can be used for global weather predictions. The models require no physical knowledge about the Earth system. In 2022, however, results improved significantly, with multiple models making legitimate claims to outperforming the skill of leading weather forecasting models with ML solutions. Leading examples are FourCastNet (Pathak et al. 2022), Pangu Weather (Bi et al. 2023), GraphCast (Lam et al. 2023) and FuXi (Chen et al. 2023). Whilst there are still caveats on quality of these models, it became of interest to weather forecasting centres worldwide to be able to run pure machine learning models in an operational context.

3.7.2 Results

In MAELSTROM, the Python library CliMetLab[4] was further developed. This package aims to ease the manipulation of weather and climate datasets for machine learning with Python. In MAELSTROM it is used to provide a standardised and easy to use interface to all the datasets provided in the project. CliMetLab provides interfaces to enable ERA5 data to be easily downloaded directly from the Climate Data Store[5] and ECMWF operational weather forecasting initial conditions from the MARS[6] archive. This lay the foundation for ai-models.

3.7.3 Further exploitation of results at ECMWF

In 2023, ECMWF created ai-models[7], a Python package to standardise the interface to these leading machine learning weather forecast models and allow them to directly use initial conditions from the ERA5 or from ECMWF's operational archive. This tool has been made open source, allowing people around the globe easier access to assess the potency of machine learning weather forecasts. At ECMWF, ai-models is used daily to run multiple external models from the open scientific research and its own recently developed machine learning forecasting system, the AIFS. The tool is run in ECMWF's operational pipeline, and the results are disseminated in chart form for free on ECMWF's website[8] (see Fig. 12 for an example). ai-models, partially thanks to developments in MAELSTROM, is accelerating the use and understanding of machine learning weather forecasting models.

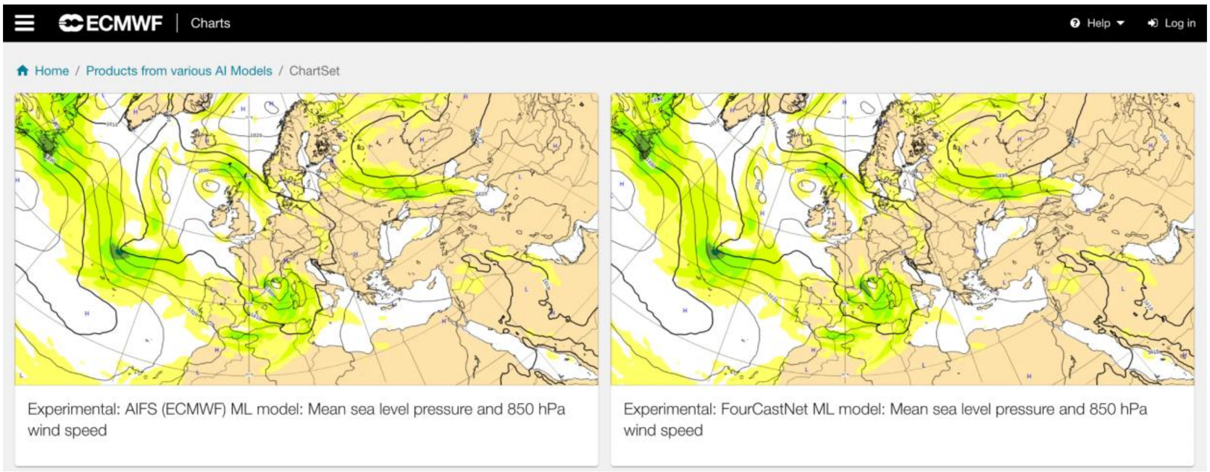
[4] https://github.com/ecmwf/climetlab

[5] https://climate.copernicus.eu/climate-data-store

[6] https://www.ecmwf.int/en/forecasts/access-forecasts/access-archive-datasets

[7] https://github.com/ecmwf-lab/ai-models

[8] https://charts.ecmwf.int/?facets=%7B%22Product%20type%22%3A%5B%22Experimental%3A%20Machine%20learning%20models%22%5D%7D



Experimental: AIFS (ECMWF) ML model: Mean sea level pressure and 850 hPa wind speed

Experimental: FourCastNet ML model: Mean sea level pressure and 850 hPa wind speed

Fig. 12: An example of charts.ecmwf.int showing the ability of users to view forecasts from data-driven machine learning models. Five different models are visible, each of them run using ai-models, which benefits from CliMetLab developments in MAELSTROM.

4 Conclusions

This report describes the work done to integrate the applications, datasets, workflow tools, and knowledge generated by MAELSTROM into the workflow at the institutes involved in the project.

By the end of the project, MAELSTROM delivered one application running operationally (AP1), one application demonstrated within an operational environment (AP3), and four production-ready applications (AP2, AP4, AP5, and AP6) ready to be exploited. MAELSTROM also contributed to the development of ai-models, a tool used operationally to run several emerging data-driven weather models (AP8).

The results from the project will have a lasting impact on the partners involved. Firstly, the benchmark datasets developed in MAELSTROM will be used by many of the partners in training staff and students, and will be used in future research projects. Project partners also have plans to create new datasets for benchmarking emerging ML models. Our work with CliMetLab will serve as a blueprint for defining these new datasets.

Secondly, the tools developed in MAELSTROM will help many of the partners build new ML-models in the future. Deep500 and other hardware and software benchmarking scripts developed in MAELSTROM will help ensure that the training pipelines for these models have high performance. The workflow tool Mantik is planned to be used at 4cast for the development and production of products for their customers.

5 References

Ashkboos, Saleh, et al., 2022: Ens-10: A dataset for post-processing ensemble weather forecasts. *Advances in Neural Information Processing Systems* **35**, 21974-21987.

Bi, K., et al., 2023: Accurate medium-range global weather forecasting with 3D neural networks. *Nature* **619**, 533-538.

Dabernig, Markus, et al., 2017: Spatial ensemble post-processing with standardized anomalies. *Quarterly Journal of the Royal Meteorological Society* 143.703. 909-916. DOI. https://doi.org/10.1002/qj.2975

Caron, Mathilde et al., 2020: "Unsupervised Learning of Visual Features by Contrasting Cluster Assignments." *ArXiv* abs/2006.09882

Chen, I., et al., 2023: FuXi: a cascade machine learning forecasting system for 15-day global weather forecast. *npj Climate and Atmospheric Science*, **6**.

Harder, Paula, et al., 2023: Hard-Constrained Deep Learning for Climate Downscaling. *Journal of Machine Learning Research,* **24**.365 (2023): 1-40.

Hatanaka, Yusuke, et al., 2023: Diffusion models for high-resolution solar forecasts. *arXiv preprint arXiv:2302.00170*.

Hess, Philipp, et al., 2023: Deep Learning for Bias-Correcting CMIP6-Class Earth System Models." *Earth's Future* **11**.10, DOI: https://doi.org/10.1029/2023EF004002

Hoehlein, Kevin, et al., 2020: A comparative study of convolutional neural network models for wind field downscaling." *Meteorological Applications,* **27**.6 e1961. DOI: https://doi.org/10.1002/met.1961

Hogan, R. J., et al., 2016: Representing 3-D cloud radiation effects in two-stream schemes: 2. Matrix formulation and broadband evaluation. Journal of Geophysical Research: *Atmospheres*, **121**, 8583-8599.

Lam, R., et al., 2023: Learning skillful medium-range global weather forecasting. *Science*, **382**, 1416-1421.

Lessig, Christian, et al., 2023: AtmoRep: A stochastic model of atmosphere dynamics using large scale representation learning. *arXiv preprint arXiv:2308.13280*

Liang, Jingyun, et al., 2021: Swinir: Image restoration using swin transformer. *Proceedings of the IEEE/CVF international conference on computer vision*.

Maas, Andrew L., et al., 2013: Rectifier nonlinearities improve neural network acoustic models. Proc. icml. **30**.

Mardani, Morteza, et al., 2024: Residual Diffusion Modeling for Km-scale Atmospheric Downscaling. *Nature Portfolio Preprint*, DOI: 10.21203/rs.3.rs-3673869/v1

Nipen, T. N., I. A. Seierstad, C. Lussana, J. Kristiansen, and Ø. Hov, 2020: Adopting Citizen Observations in Operational Weather Prediction. Bull. Amer. Meteor. Soc., **101**, E43–E57, https://doi.org/10.1175/BAMS-D-18-0237.1.

Pathak, J., et al., 2022: FourCastNet: A Global Data-driven High-resolution Weather Model using Adaptive Fourier Neural Operators. *arXiv*. https://arxiv.org/abs/2202.11214.

Price, I., and Rasp, S, 2022: Increasing the accuracy and resolution of precipitation forecasts using deep generative models. *International conference on artificial intelligence and statistics*. PMLR.

Rasp, Stephan, et al., 2023: Weatherbench 2: A benchmark for the next generation of data-driven global weather models, *arXiv preprint arXiv:2308.15560*.

Ramachandran, P. et al., 2017: Searching for activation functions. *arXiv preprint arXiv:1710.05941*.

Scherrer, S. C., 2020: Temperature monitoring in mountain regions using reanalyses: lessons from the Alps. *Environmental Research Letters*, ***15.***4, 044005. DOI: 10.1088/1748-9326/ab702d

Sha, Yingkai, et al., 2020: Deep-learning-based gridded downscaling of surface meteorological variables in complex terrain. Part I: Daily maximum and minimum 2-m temperature. *Journal of Applied Meteorology and Climatology,* **59**.12: 2057-2073.

Shi, W., et al., 2016: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. Proceedings of the IEEE conference on computer vision and pattern recognition. DOI: http://dx.doi.org/10.1109/CVPR.2016.207

Ukkonen, P., and R. J. Hogan., 2024: Twelve Times Faster yet Accurate: A New State-Of-The-Art in Radiation Schemes via Performance and Spectral Optimization. *Journal of Advances in Modeling Earth Systems*, **16**, e2023MS003932.

Vosper, Emily, et al., 2023: Deep Learning for Downscaling Tropical Cyclone Rainfall to Hazard-Relevant Spatial Scales. *Journal of Geophysical Research: Atmospheres* **128**.10: e2022JD038163. DOI: https://doi.org/10.1029/2022JD038163

Document History

Version	Author(s)	Date	Changes
0.1	Thomas Nipen et al.	21/03/2024	Version for review
1.0	Thomas Nipen et al.	27/03/2024	Final version

Internal Review History

Internal Reviewers	Date	Comments
Peter Dueben (ECMWF)	23/03/2024	Minor comments and suggestions provided
Mats Brorsson (UL-SnT)	25/03/2024	Minor comments and suggestions provided

Mattia Paladino (E4)	25/03/2024	Minor comments, request to make Fig 5 readable and to add the chapter of the conclusions

Estimated Effort Contribution per Partner

Partner	Effort
MetNor	1pm
Total	**1pm**

This publication reflects the views only of the author, and the European High-Performance Computing Joint Undertaking or Commission cannot be held responsible for any use which may be made of the information contained therein.