Technical Memo

European Centre for Medium-Range Weather Forecasts

878

Machine learning at ECMWF: A roadmap for the next 10 years

Peter Dueben, Umberto Modigliani, Alan Geer, Stephan Siemen, Florian Pappenberger, Peter Bauer, Andy Brown, Martin Palkovič, Baudouin Raoult, Nils Wedi, Vasileios Baousis

January 2021

Series: ECMWF Technical Memoranda

A full list of ECMWF Publications can be found on our website under:
http://www.ecmwf.int/en/publications

Contact: library@ecmwf.int

Executive summary

During the last decade, artificial intelligence (AI), machine learning, and data volume have developed at an unprecedented pace, and it is now evident that many scientific disciplines will need to revise their work modes to become more data centric in order to make the most out of these developments. AI and machine learning offer great opportunities throughout the workflow of numerical weather prediction (NWP) and climate services, and the science community is currently exploring how the new capabilities of AI and machine learning will change the future of Earth system science. First results show great potential.

However, the scope and speed of developments also generate challenges for weather and climate modelling centres such as ECMWF, in particular regarding the necessary know-how that needs to be established, the software and hardware infrastructure that needs to be developed, and the integration of machine learning and conventional tools within the prediction workflow. These challenges need to be addressed within a comparably short period of time to keep up with changing needs of the weather and climate modelling community and ECMWF's Member and Co-operating States. This document therefore sets out a roadmap for the next ten years that identifies the challenges, provides potential solutions, and defines steps to channel the many distributed science and technology projects that study machine learning for weather and climate predictions into a coordinated effort. While the roadmap does not provide a scientific workplan for machine learning activities, due to the number and diversity of the application areas, it outlines the path towards more coordinated solutions for the challenges ahead, and to generate synergies between the different machine learning efforts.

What, why and how

What are AI and machine learning?

AI is intelligence demonstrated by machines, unlike the natural intelligence displayed by humans. Machine learning is the study of computer algorithms that improve automatically through learning from data without being explicitly programmed. Machine learning represents the most relevant subset of AI for Earth system science. The range of methods that can be counted as machine learning is wide, but machine learning can mainly be classified into two groups:

1) *Supervised* methods learn to represent a certain task based on labelled data. This task can, for example, perform classification (there is a tropical cyclone visible in this pressure field) or regression (the expected daily precipitation for London tomorrow is 2.3 mm).
2) *Unsupervised* methods learn to distinguish data samples based on unlabelled data, for example via clustering and dimensionality reduction tasks. Most machine learning methods that are currently being investigated for Earth system science are supervised as they are easier to configure and interpret. However, unsupervised methods are receiving a growing amount of attention.

Why now?

AI and machine learning became increasingly popular and have achieved human-level performance in many challenging application areas such as image and speech recognition, gaming, finance and many more. This development has been fuelled by:

1) An unprecedented increase in data volume which makes it more-and-more challenging for scientists to extract all relevant information using conventional methods. The business-as-usual regarding data handling and information will not allow scientists to cope with the hundreds of terabytes of data that will likely be produced within a single day by operational weather predictions in the near future.
2) An increase of knowledge behind AI and machine learning, with more than 100 papers published every day, that allows the development of machine learning applications that are customised towards the needs of specific applications.
3) Developments in computing hardware to allow for the training of machine learning tools with billions of trainable parameters from many terabytes of data.
4) Freely available, open-source software frameworks that are easy to use (e.g. TensorFlow and PyTorch) and allow the development of complex machine learning applications based on a couple of hundred lines of Python code.

The use of some machine learning algorithms and statistical models is well established in the NWP community, for example via the use of principle component analysis or the use of data assimilation techniques that can also be interpreted as machine learning (Bocquet et al., 2020; Geer, 2021). However, regarding the use of complex machine learning techniques such as deep neural networks, Earth system science is still lagging behind other research disciplines.

How will AI and machine learning change NWP and climate services?

As the Earth system is complex with non-linear behaviour and as there are hundreds of petabytes of data about the Earth system available, including both observations and model output, machine learning provides a very powerful toolbox to improve weather and climate predictions. Machine learning can be used to improve the computational efficiency of weather and climate models, to extract information from data, or to post-process model output, in particular if data-driven machine learning methods can be combined with conventional tools. For machine learning, a wide range of potential application areas show great potential throughout the workflow of NWP and climate services and throughout ECMWF (see Figure 1). Regarding post-processing, ECMWF will focus on supporting the Member and Co-operating States to develop machine learning tools and the use of machine learning for diagnostic purposes, for example to understand the dynamics and weaknesses of predictions.

The number of successful use cases for sophisticated machine learning techniques at ECMWF is growing quickly – such as deep neural networks and decision trees. Early success stories across ECMWF's workflow include the use of neural networks for SMOS soil moisture data assimilation for the land surface (Rodríguez-Fernández et al., 2019) and the use of neural networks within the weak-constraint 4D-Var framework (Bonavita and Laloyaux, 2020). Deep learning has been used successfully for the emulation of the gravity wave drag parametrization schemes (Chantry et al., 2021) and the deep learning emulators could be used to generate tangent linear and adjoint model code for 4D-Var data assimilation (Hatfield et al., 2021). Furthermore, decision trees have been used for the post-processing of ensemble predictions for precipitation (Hewson and Pillosu, 2020), and a study on the use of machine learning for anomaly detections has been performed for the logs of ECMWF's data servers to detect and predict system failures via a project of the ECMWF Summer of Weather Code 2020[1].

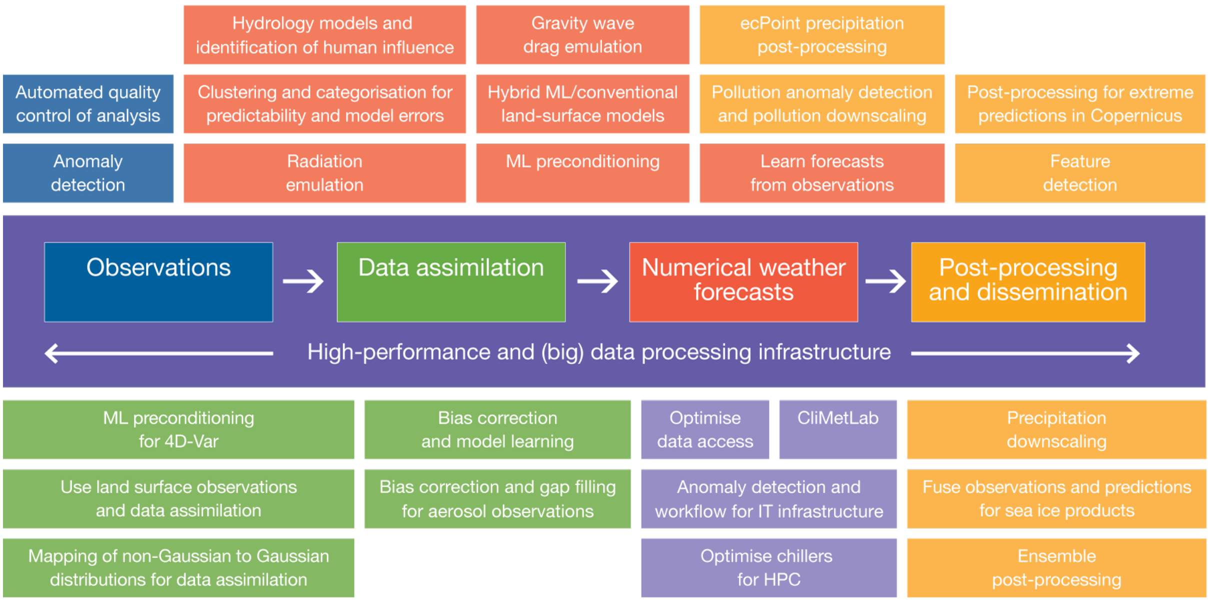

Figure 1: Machine learning applications at ECMWF that are already being explored or planned. The colour-coding of the boxes corresponds to the respective component of the workflow for NWP

[1] https://esowc.ecmwf.int/

Challenges for the adoption and effective use of AI and machine learning at ECMWF and how they will be addressed

This section discusses **challenges** and the **approaches** ECMWF will take to address them.

Machine learning scientists and Earth Science domain scientists often follow a different philosophy. The former tend to solve data science problems that optimise for specific goal functions (e.g. the reduction of root mean square error of precipitation at 48-hour lead time), while the latter aim to improve and validate models with physical understanding and checks for physical consistency (e.g. conservation laws or process feedbacks). Domain scientists are sometimes defensive regarding the establishment of machine learning as they consider the new capabilities as a threat rather than as an extension of their toolbox. This makes communication and collaboration difficult, but it also represents a barrier for the uptake of machine learning solutions by domain scientists who do not trust black box approaches that do not offer physical understanding. This is in particular the case as many of the machine learning architectures that are currently applied have been developed for other domains such as image recognition and do not allow for the introduction of domain knowledge into solution design. There is a risk that solutions for specific applications will be developed in parallel with no synergies between domain scientists on the one side and machine learning scientists on the other.

Approach: *Overcoming this barrier requires close collaborations between machine learning scientists and domain scientists to develop physically consistent machine learning solutions for operational use that exploit the full potential of the new toolbox of advanced machine learning and complement existing physically based solutions. Explainable AI and physics-informed machine learning, which try to blend machine learning with physical knowledge to achieve solutions that are physically more consistent, will be explored further (McGovern et al., 2019; Reichstein et al., 2019). Furthermore, trustworthy AI will be explored to improve our understanding of how machine learning methods are working and to shed some light into the black box.*

While some of the applications from machine learning in Earth system sciences are conceptually close to the use of machine learning methods in other domains (such as the detection of tropical cyclones in model output which can be formulated as an image recognition task), many **will require customised machine learning solutions**. Physical fields, for example, are often stored on unstructured grids on the sphere which do not allow for a simple application of convolutions in space or time which are a core element of many machine learning approaches. While the vertical dimension of atmosphere models is structured, the physical fields still show very different dynamics close to the surface and at the model top, which will, again, make it difficult to apply standard convolution approaches. Furthermore, physical fields need to follow physical constraints such as conservation laws or a limit to positive values (e.g. for precipitation).

Approach: *Customised machine learning solutions for domain-specific problems (such as the capability to perform convolution operations in neural networks on unstructured grids on the sphere) will need to be developed. The customised solutions can then be applied to many different machine learning applications within the domain and serve as benchmark solutions. The quickest path to customised solutions is the development of benchmark datasets and problems – including datasets, cost functions, and example solutions – that allow machine learning scientists from different groups and institutions to make quantitative comparisons of machine learning solutions (e.g. WeatherBench in Rasp et al., 2020).*

Machine learning tools should not be used only to emulate or accelerate model components but also to *improve* the models. This will most often require training machine learning tools from Earth system observations and, therefore, comparing model trajectories to observations that represent the same physical situation in space and time. However, it is **difficult to learn from Earth system observations** as they are sparse, irregular, and uncertain, and extracted from heterogeneous instrumentation (including satellite radiances) which typically cannot be compared to model fields directly.

Approach: *The best way to relate model simulations to observations of the Earth system is data assimilation. Machine learning approaches and data assimilation have much in common and machine learning for Earth system science should therefore adapt to the data assimilation workflow in many instances (Geer, 2021). Examples include the use of observation errors to represent varying levels of uncertainty, observation operators to map from regular model grids to irregular, sparse observations, and the use of physical model components or layers to impose physical constraints on otherwise machine-learned networks. On the other hand, as visible in Figure 1, there are already many interesting applications for machine learning to work with structured datasets to improve the processing of observations (for example with observation operators) and data assimilation (for example via the learning of model or observation bias).*

Current efforts to improve the mapping from satellite measurements to surface maps (often also based on machine learning) need to be followed closely as they offer new opportunities to improve land surface parametrizations and can serve as a reference truth for atmospheric dynamics close to the surface. Furthermore, machine learning will very likely be essential to extract relevant information from Internet of Things (IoT) data and other data sources such as traffic counts, energy production and transport analytics to supplement current Earth observations. IoT data is typically noisy but available in very large amounts, and therefore difficult to handle using conventional methods.

The training of machine learning tools requires **data**. As the complexity of machine learning tools can be increased arbitrarily, the only limits for the accuracy of machine learning methods are the amount of data that is available for training and the limits of the computational and data handling infrastructure. More data allow more sophisticated machine learning solutions to be designed. Machine learning users will therefore show a different and more greedy data access pattern when compared to conventional users, towards larger and more selective data access (e.g. retrieving a single field over a specific area of the globe for a long period of time). This generates a significant challenge for data centres such as ECMWF as data production and use is already growing fast for the conventional workflow.

Approach: *The computing infrastructure at ECMWF needs to be prepared for High Performance Data Analytics (HPDA) and research which is increasingly data driven. This requires a serious effort to explore the capabilities of heterogeneous hardware that will be available for future high-performance computing (HPC) to reduce the I/O bottleneck when handling large amounts of data. To buffer the increase in data requirements due to machine learning, the data workflow will need to be organised in a way that allows easy access to the most prominent fields and data products and that takes the heterogeneity of hardware options for data storage and access into account (e.g. tapes vs. discs). Data access patterns will need to be anticipated, and therefore this will require the involvement of the machine learning community. The generation of benchmark datasets which can cover a large fraction of user requests as well as the communication of existing datasets that have already been assembled should further reduce the need for individual scientists to assemble large data on their own.*

When compared to conventional methods in NWP and climate services, machine learning requires a **different set of tools regarding both software and hardware**. Most Earth system models are still based on Fortran code and are typically run on CPU-based supercomputing hardware. On the other hand, machine learning tools are typically based on Python code and Python libraries as well as Jupyter notebooks and are trained and used most efficiently on GPU hardware. Most of the computational cost for supervised learning is generated by the training of the machine learning tool, while the application of the tool (the "inference") is typically very cost efficient. As part of the code refactorization to improve portability, models are now being rewritten into domain-specific languages and in some cases also Python or Julia code (Bauer et al., 2020), including the Finite Volume version of the ECMWF Integrated Forecasting System (IFS-FVM). However, it will still take a couple of years until these developments reach a large fraction of domain scientists.

Approach: *Training is required to support domain scientists at ECMWF to start working with machine learning tools and facilitate a smooth start in new software environments. Domain scientists need to be supported with efficient tools and customised solutions to make the first steps in the new environment easier (e.g. to read GRIB or netCDF data into Python). Trends in the fast-moving area of machine learning software need to be monitored and solutions need to be adapted.*

To allow the efficient training and application of machine learning, hardware that is suitable for machine learning must also be available instead of standard CPU-based hardware optimised for conventional applications. It will also require relevant machine learning software to be installed on all computer hardware, from desktops to supercomputers.

As software and hardware requirements are different, solutions are also required to **integrate machine learning tools into the conventional NWP and climate services workflow**. It is, for example, difficult to introduce machine learning tools written in Python into the Fortran code of the Integrated Forecasting System (IFS), and there is still no experience of how to update machine learning tools that need adjustments when preparing new model cycles for operational use.

Approach: *To reduce the overall workload, centralised software solutions are needed to integrate machine learning and conventional tools within ECMWF's workflow. Domain scientists need to be supported when applying the centralised solutions, and the solutions need to be aligned with the efforts of the ECMWF Scalability Project, regarding model portability.*

Machine learning is **a new skill that needs to be developed and established** at institutions such as ECMWF. While machine learning experience is still limited, it is also spread across the entire workflow of ECMWF, which makes it challenging to generate synergies and knowledge exchange between the pioneering scientists involved when applying new methods (such as complex decision trees) in a different context (such as wildfires or parametrization scheme emulation). Furthermore, machine learning solutions are still fragile as they depend on the expertise of individual scientists or external collaborators who have developed the solutions. This makes it difficult to guarantee the level of reproducibility required for operational weather predictions and climate services.

Approach: *To allow for synergies between different machine learning efforts and to guarantee the reproducibility of solutions, a team of experts is needed to guide and support individual scientists pursuing their machine learning approaches, and to organise centralised software solutions. The team needs to make an effort to identify needs, and to communicate ongoing efforts to address those needs. It needs to show high availability to individuals, flexibility to address individual challenges but respect for the existing organisational structure.*

First steps done

ECMWF has engaged in a number of collaborations with external partners – and in particular machine learning specialists – to explore the potential of machine learning for applications throughout the NWP workflow (see table in the appendix). ECMWF scientists have given invited talks at a number of international AI conferences, and a number of machine learning publications are in the pipeline for publication or have already been published. ECMWF has established the role of an AI and machine learning coordinator to orchestrate the pan-institutional efforts and has also started to acquire significant GPU hardware that is suitable for machine learning projects for both the new HPC and the European Weather Cloud[2] which is being developed in collaboration with EUMETSAT. As many of the interactive developments of machine learning tools are performed on cloud hardware that allows interactive developments with Python, Jupyter notebooks or Julia and the use of high-end hardware in a scalable way, the European Weather Cloud will be an important resource for machine learning training and applications in the future. The first scientists are already using it for machine learning on large datasets.

The first benchmark dataset for machine learning applications in weather and climate modelling has been published with contributions from ECMWF (Weatherbench; Rasp et al., 2020). Further efforts with ECMWF contributions are in the making (including a project at the World Meteorological Organization (WMO) on subseasonal-to-seasonal (S2S) predictions, a post-processing initiative coordinated by EUMETNET, and land-surface modelling within the GEWEX framework).

Next to existing activities for the use of Python at ECMWF (for example via training and data APIs), an initiative called the CliMetLab[3] has been launched that is specifically targeted to support machine learning applications to simplify access to climate and meteorological datasets. CliMetLab is including data import from the ECMWF Meteorological Archival and Retrieval System (MARS) and the Copernicus Climate Change Service Climate Data Store (CDS) into Python environments and allows users to focus on science instead of technical issues such as data access and data formats.

To foster collaborations and to share experience between domain scientists and Member and Co-operating States, ECMWF has organised two conferences on machine learning – namely the 1st Artificial Intelligence for Copernicus Workshop[4] and the ECMWF-ESA Workshop on Machine Learning for Earth System Observation and Prediction[5]. Furthermore, a machine learning seminar series took place in 2020 for which talks were live-streamed and recorded[6]. One advanced and four introductory training courses for ECMWF staff were organised in 2020 to establish know-how on machine learning within the institute.

[2] https://www.europeanweather.cloud/
[3] http://climetlab.readthedocs.io/
[4] https://atmosphere.copernicus.eu/1st-artificial-intelligence-copernicus-workshop
[5] https://events.ecmwf.int/event/172/
[6] https://www.ecmwf.int/en/learning/workshops/machine-learning-seminar-series

The move to an open data policy for historical information in ECMWF's data repository[7] will open further possibilities for collaborations with external machine learning experts in the future. The new *Center of Excellence in Weather & Climate Modelling* between ATOS and ECMWF[8] – with support from AMD, Mellanox, Nvidia and DDN – also includes a project that is dedicated to developments in machine learning. The project will develop customised machine learning solutions that are optimised for use in the vertical direction and on the unstructured horizontal grid of the IFS. Furthermore, the project will support the efficient integration of the machine learning solutions into the conventional HPC workflow of weather predictions and climate services at ECMWF on the supercomputer.

ECMWF has also been successfully contributing to externally funded projects. This includes the MAELSTROM project that is coordinated by ECMWF and that has been funded under EuroHPC-JU. MAELSTROM will perform a co-design cycle to develop benchmark datasets, vanilla machine learning solutions, software frameworks, and hardware system designs that are customised for machine learning applications in Earth system science. The know-how and infrastructure that will be developed in MAELSTROM will be available for adaptation at ECMWF. ECMWF is also contributing to the AI4Copernicus project that has been funded under H2020-ICT to develop a software infrastructure for machine learning applications on the Copernicus Data and Information Access Services (DIAS). Furthermore, ECMWF is a partner in the CLINT H2020-LC project and will investigate the use of machine learning to identify key aspects of the three-dimensional atmospheric structure preceding the genesis of tropical cyclones, such as tropical waves in the atmosphere and ocean heat structure.

How to progress – the big picture

ECMWF aims to enable the ECMWF Member and Co-operating States and the weather and climate modelling community in Europe to make the most of machine learning in the years to come, and to show how machine learning fits into, benefits or replaces existing core developments to improve NWP and climate services. To fulfil this aim, ECMWF will continue to address five main objectives:

Objective 1	Objective 2	Objective 3	Objective 4	Objective 5
Explore machine learning applications across the weather and climate prediction workflow and apply them to improve model efficiency and prediction quality.	Expand software and hardware infrastructure for machine learning.	Foster collaborations between domain and machine learning experts with the vision of merging the two communities.	Develop customised machine learning solutions for Earth system sciences that can be applied to various applications and at scale on current and future supercomputing infrastructure.	Train staff and Member and Co-operating State users and organise scientific meetings and workshops.

However, ECMWF will also identify limits of state-of-the-art machine learning for Earth system modelling, for example regarding the representation of non-linear systems and physical consistency with black box approaches and application areas for which machine learning approaches will *not* beat existing solutions (including some of the application areas named in Figure 1).

[7] https://www.ecmwf.int/en/about/media-centre/news/2020/ecmwf-moves-towards-policy-open-data
[8] https://www.ecmwf.int/en/about/media-centre/news/2020/ecmwf-and-atos-launch-center-excellence-weather-climate-modelling

As the exploration of sophisticated machine learning tools for weather and climate modelling is still at an early stage, ECMWF will foster **scientific studies** of machine learning methods in applications which are meaningful for Earth system science but, at the same time, small enough to allow detailed and quantitative comparisons between different machine learning solutions. Manageable problems in terms of data use and the complexity of machine learning tools will allow for fast progress when exploring physics-informed machine learning and trustworthy AI, and hybrid modelling approaches that combine conventional and machine learning tools. Small problems will also help when exploring uncertainty quantification and uncertainty representation, and for the development of customised machine learning solutions for domain-specific problems, for example the use of Graph Neural Networks to perform convolutions on unstructured model grids on the sphere.

At the same time, **large-scale machine learning solutions** are being tested and developed with many millions of trainable parameters that are capable of taking the three-dimensional state of the global atmosphere as input, train from many terabytes of data, and require the use of supercomputers. This will be necessary to explore the limits and potential of the new tools within Earth system modelling and to be prepared for large-scale machine learning applications in the future, in particular as machine learning has a fundamental influence on future developments of HPC infrastructure.

How to progress – specific milestones

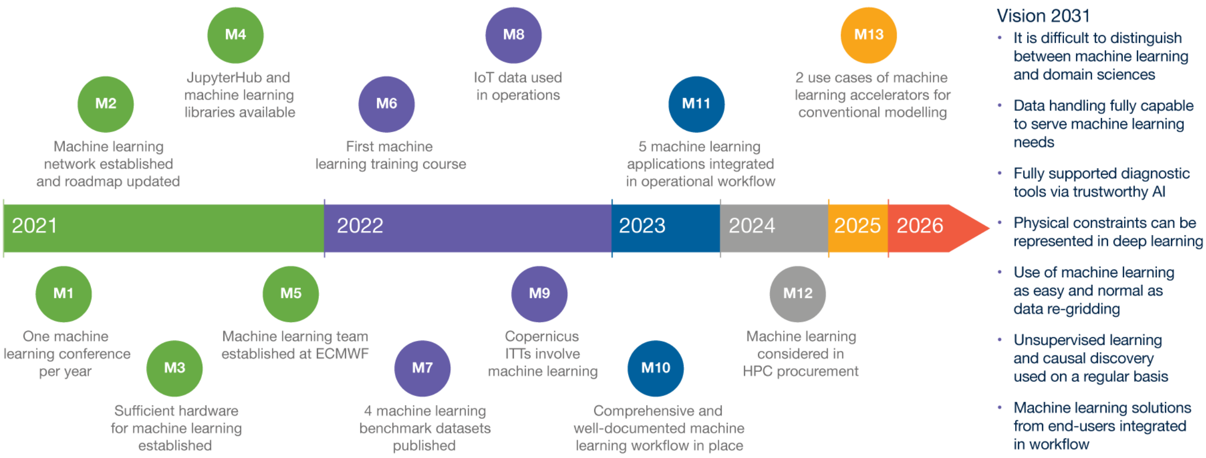

Figure 2: Timeline of machine learning developments at ECMWF with all milestones.

For the next five years, we have defined milestones – defined at quarter (Q) or half (H) years – for the technical and organisational support of machine learning activities at ECMWF and within the Member and Co-operating States. We also provide a vision for the use of machine learning by 2031. Figure 2 provides a timeline for the developments.

<u>The next 5 years</u>

Milestone 1, Q1, 2021: At least one conference that is focusing on machine learning is organised by ECMWF every year.

Milestone 2, Q2, 2021: A network for collaborations between machine learning experts in the Member and Co-operating States and the national meteorological services has been established and the machine learning roadmap and milestones have been refreshed based on Member and Co-operating State feedback.

Milestone 3, Q3, 2021: A sufficient amount of hardware that allows the efficient training and inference of machine learning tools (such as GPUs) has been established. New hardware technologies for machine learning will be explored.

Milestone 4, Q4, 2021: A JupyterHub and machine learning libraries are accessible on key computing hardware at ECMWF.

Milestone 5, Q4, 2021: A team for machine learning has been established which is distributed across the organisation and covers key elements of the machine learning value chain.

Milestone 6, Q1, 2022: A machine learning training course for Member and Co-operating State users is established.

Milestone 7, Q2, 2022: At least four machine learning benchmark datasets are published.

Milestone 8, Q3, 2022: Machine learning tools are used for quality control and to design observation operators for IoT data to supplement current Earth observations within the NWP and climate service workflow.

Milestone 9, Q4, 2022: ITTs of the next phase of the Copernicus programme involve machine learning.

Milestone 10, H1, 2023: An efficient and well-documented centralised machine learning workflow is established at ECMWF that covers data retrieval, data pre-processing, machine learning training, solution evaluation and inference within the IFS.

Milestone 11, H2, 2023: At least five machine learning applications are integrated into the operational workflow.

Milestone 12, 2024: Machine learning applications are considered as benchmarks in the HPC procurement.

Milestone 13, 2025: At least two use cases for machine learning accelerators to improve computational efficiency of conventional model components are realised within operational predictions.

Scientific milestones are not included in the list above as it would not be credible to quantify breakthroughs as machine learning solutions will be integrated with conventional tools, and as a discussion of individual applications would be beyond the scope of this short roadmap document due to the large number of application areas within ECMWF. However, a list of the ongoing scientific projects and an outline of future steps is provided in the appendix (see also Figure 1). The machine learning areas that appear most promising for implementation in the operational workflow within the next three years (see Milestone 11) are the processing of observations (see SMOS project) and observation operators (aim 2022), bias correction in data assimilation (Bonavita and Laloyaux, 2020, aim 2022), the emulation of physical parametrization schemes with ongoing efforts regarding the gravity wave drag and radiation that include the generation of tangent linear and adjoint model code (aim 2023), post-processing of ensemble predictions (Baran et al., 2020; Hewson and Pillosu, 2020; Groenquist et al., 2020, already used), or the scheduling of jobs or detection of anomalies on the HPC system. These machine learning areas will be tested and pushed to operational use if results are convincing. If not, other applications will be considered.

Additionally, we expect that an efficient solution to couple Fortran code and machine learning libraries within the IFS will be made available towards the end of 2021. We also expect that generic solutions for machine learning applications in the vertical direction of IFS and for three-dimensional applications on the unstructured cubic octahedral reduced Gaussian grids (as used within IFS) will be available by 2023.

The long-term vision for 2031

We anticipate that it will be increasingly difficult to distinguish between scientists working on machine learning and domain scientists in the future and that it will no longer be possible to identify the tools that were originally targeted for machine learning applications in ten years from now. Our vision is that by 2031, machine learning is fully integrated into NWP and climate services and has improved predictions and the use of predictions in many areas of the workflow. Special requirements for data retrievals for machine learning applications are well understood and the data handling has been adjusted to fit those needs and to provide all users in a user group with the required data with only limited duplication of data requests. Customised machine learning solutions have been developed for a number of application areas in weather and climate modelling that serve as blueprints for new machine learning applications in the domain. Furthermore, diagnostic tools that are based on trustworthy AI have been developed for Earth system scientists to explore and understand the functionality of sophisticated machine learning solutions. It is understood how to incorporate physical constraints, such as conservation laws, into neural network design and training. Eventually, the use of sophisticated machine learning tools is as easy and normal for relevant domain scientists as the re-gridding of data to grids with different resolutions. Not only supervised learning, but also unsupervised learning and causal discovery methods are used on a regular basis. Finally, machine learning solutions from end-users can be integrated into the NWP and climate services workflow at ECMWF to avoid heavy data processing and to allow for interactive use.

Closing remarks

Following the steps outlined in this roadmap will enable ECMWF to prepare for evolving needs of scientists and analysts towards a more data-driven workflow and to support the Member and Co-operating States to make the most of new capabilities of machine learning as soon as possible.

The scope of the roadmap will be adjusted depending on future developments of the EU's *Destination Earth* initiative that has AI and machine learning as one of the main building blocks to develop Digital Twins of the Earth system. The *Digital Twin on Weather-induced and Geophysical Extremes* shows a particular need for machine learning applications to help improve model efficiency (in particular via the use of machine learning preconditioners for linear solvers or the emulation of model components with neural networks), enhance the quality of local predictions (for example via local down-scaling, bias-correction and uncertainty quantification), and introduce customised, interactive applications from end-users into the prediction workflow (for example via the automatic detection of features during simulations). For the *Digital Twin on Climate Change Adaptation*, machine learning will enable a more efficient extraction of information from large datasets or an understanding of causality and physical connectivity via unsupervised learning.

Machine learning efforts at ECMWF will also be aligned with current efforts from ESA to use Earth observations to improve global maps that can be used for modelling and, potentially, augment ECMWF's data assimilation efforts. These maps will facilitate the development of better land surface parametrizations and the evaluation of the required complexity of these parametrizations, for example for the development of an urban tile in global simulations.

Appendix

The table below provides a list of scientific projects based on machine learning that are currently being investigated at ECMWF.

Project	Status	Next steps
Anomaly detection in system logs of data servers	First successful tests performed as an ECMWF Summer of Weather Code challenge	Uptake of software solutions and more detailed testing at ECMWF
Benchmarking of ensemble post-processing methods (EUMETNET project)	Review paper published in BAMS (Vannitsem et al. 2020)	Development phase on the European Weather Cloud
Bias correction learned from data assimilation using deep learning	First paper published (Bonavita and Laloyaux 2020) and first tests with bias correction implemented in the forecast model performed	Move to learning from three-dimensional inputs on unstructured grids and continue evaluation of bias correction
Data-driven transport module for atmospheric tracers	Research proposal submitted	Gather training dataset and start the work
Data monitoring via machine learning to increase the robustness of anomaly detection for observations and to highlight model/DA weakness with a consistent signature on model-observation departures	A prototype of the alarm system based on random forest classifiers has been tested	Work will continue throughout 2021
Emulation of the gravity wave drag parametrization schemes using neural networks	Paper in preparation in collaboration with the University of Oxford	Extend to other parametrization schemes
Emulation of the radiation scheme using neural networks	First stable results with neural network emulator performed in the IFS in a collaboration with NVIDIA. However, results are not neutral yet	Continue training of emulators from new dataset as part of the MAELSTROM project
Emulation of 3D cloud effects from the SPARTACUS radiation scheme using neural networks	First offline tests in a collaboration with Reading University are successful	Perform tests of the impact of the correction within IFS simulations

Learning earth system models from observations: machine learning or data assimilation?	Paper describing similarities between machine learning and DA in a Bayesian framework, suggesting ways to combine the two (Geer, 2020). Similarities also emphasised in Boukabara et al. (2020).	Practically demonstrate ways that data assimilation ideas (e.g. observation errors, observation operators, physical constraint layers) can help improve typical machine learning applications. In parallel, continue working on parameter estimation for cloud and precipitation physical assumptions
Improving flood forecast skill using post-processing along river network under ungauged reaches	Prototype code under way. Start of evaluation skill improvement of operational EFAS method (only valid for gauged catchments)	Implement new method and benchmark results against EFAS operational
Investigate the potential of causal inference/discovery methods for predictability research and model evaluation	Considering extending the TIGRAMITE software to work seamlessly with initialised ensemble forecasts	Test TIGRAMITE with sub-seasonal and seasonal case studies
Machine learning for satellite bias correction	Needs to be finalised	Consider using the newly developed SSMIS bias correction in the operational system
ML interpretability techniques to better understand (conditional) forecast errors and how to correct for them	Work only just started with a literature review	Application to various datasets (e.g. observation supersite)
Machine learning technique to identify irrigated areas	Work only just started with a literature review	Define Earth observation data attributes and explore different methods using SMOS soil moisture
Post-processing of precipitation from ensemble simulations	Paper published (Hewson and Pillosu, 2020) and operational implementation in place	Further tests with deep learning and other mapping procedures refining calibration software; extending post-processing to 2m temperature; applying to ERA5 and extended-range forecasts
Post-processing of ensemble predictions for T850 and Z500 with deep learning	Paper published in a collaboration with ETH Zurich (Groenquist et al., 2020)	Further tests with more complex input states, more ensemble members and unstructured grids

Sea ice emissivity and fraction	Project starting to implement a sea-ice emissivity and fraction analysis in 4D-Var based on microwave satellite radiances, using an augmented control vector and an empirical model	Evaluate whether machine learning methods can do better than traditional empirical approaches in building an empirical model of sea ice emissivity from satellite data and model fields
Tropical cyclone feature detection in forecasts	One project from the ESOWC 2020 and one project in collaboration with NOAA. An initial validation of the NOAA model on ERA5 data from CDS has started (30-year dataset, model trained across 14 GPU's)	Benchmarking the software of ESOWC 2020, and design and development of a pre-operational machine learning based cyclone detection service from the project with NOAA
Use of machine learning to predict fire ignition occurrences from lightning forecasts	Paper in press in MetApp (Coughlan et al., 2020)	Use of machine learning to derive fuel availability from vegetation indices
Use neural networks to retrieve soil moisture from SMOS and ASCAT observations	Two operational products produced in NRT for SMOS. One is delivered to ESA and the other is assimilated into ECMWF land-surface assimilation system via the SEKF. Initial research into similar approach for ASCAT (Aires et al., 2020, in review).	Re-training of SMOS neural networks and further validation of the ASCAT neural networks approach
Use machine learning with SMOS soil moisture data to predict river discharge categories	First exploration finished showing promising theoretical predictability when driven with observed weather input.	New tests under way to benchmark results with the EFAS hydrological prediction
Use of neural network emulators to generate tangent linear and adjoint model code for 4D-Var	First successful analysis experiments have been performed for the gravity wave drag emulator in collaboration with University of Oxford	Submit paper and perform further tests with other parametrization schemes
Use of neural networks for preconditioning of linear solvers for atmosphere models	First paper online (Ackmann et al., 2020) for study with shallow water model in collaboration with the University of Oxford and NCAR	Write journal paper and increase complexity of the application

Use of neural networks for forecast simulations	First paper (Dueben and Bauer, 2018) and benchmark dataset published (Rasp et al., 2020)	Explore use of unstructured grids in a collaboration with EPFL
Weather normalisation of pollution changes in observations using gradient boosting	Paper in review in ACP (Barre et al., 2020) showing the need for this normalisation during the COVID lockdowns over Europe to isolate the contribution of emission changes	Use routinely and globally. Possible collaboration with IPSL to use machine learning to perform source inversion for air quality.

References

Ackmann, J., P. D. Dueben, T. Palmer, & P. Smolarkiewicz, 2020: Machine-Learned Preconditioners for Linear Solvers in Geophysical Fluid Flows. https://arxiv.org/abs/2010.02866.

Aires et al, 2021: Statistical approaches to assimilate ASCAT soil moisture information: Part I Methodologies and first assessment. *Q.J.R. Meteorol. Soc.* In review.

Baran, S, Baran, Á, Pappenberger, F, & Ben Bouallègue, Z., 2020: Statistical post- processing of heat index ensemble forecasts: Is there a royal road? *Q.J.R. Meteorol. Soc.*, 146:3416-3434, https://doi.org/10.1002/qj.3853.

Barré, J. et al., 2020: Estimating lockdown induced European NO2 changes. *Atmos. Chem. Phys. Discuss.*, https://doi.org/10.5194/acp-2020-995.

Bauer, P. et al., 2020: The ECMWF Scalability Programme: Progress and Plans. *ECMWF Technical Memorandum*, **857**, doi: 10.21957/gdit22ulm.

Bonavita, M. & P. Laloyaux, 2020: Machine Learning for Model Error Inference and Correction. *Earth and Space Science Open Archive*, 10.1002/essoar.10503695.1.

Bocquet, M., J. Brajard, A. Carrassi & L. Bertino, 2020: Bayesian inference of chaotic dynamics by merging data assimilation, machine learning and expectation-maximization. *Foundations of Data Science*, **2(1)**, pp. 55-80.

Boukabara, S. & Coauthors, 2020: Outlook for Exploiting Artificial Intelligence in the Earth and Environmental Sciences. *Bull. Am. Meteorol. Soc.*, doi: https://doi.org/10.1175/BAMS-D-20-0031.1.

Chantry, M., S. Hatfield, P. Dueben, I. Polichtchouk & T. Palmer, 2021: Machine learning emulation of gravity wave drag in numerical weather forecasting. In preparation.

Coughlan et al., 2021: Using machine learning to predict fire ignition occurrences from lightning forecasts. Met App in press.

Geer, A. J., 2021: Learning earth system models from observations: Machine learning or data assimilation? *Phil. Trans. A*. In press (preprint: https:/doi.org/10.21957/7fyj2811r).

Grönquist, Peter, et al., 2020: Deep Learning for Post-Processing Ensemble Weather Forecasts. *Phil. Trans. A*.

Hatfield, S. M. Chantry, P. Dueben, P. Lopez, A. Geer, I. Polichtchouk & **T. Palmer**, 2021: Neural networks as the building blocks for tangent-linear and adjoint models. In preparation.

Hewson, T. & **F. Pillosu**, 2020: A new low-cost technique improves weather forecasts across the world. arXiv:2003.14397v1.

McGovern, A., R. Lagerquist, D. John Gagne, G. E. Jergensen, K. L. Elmore, C. R. Homeyer & **T. Smith**, 2019: Making the Black Box More Transparent: Understanding the Physical Implications of Machine Learning. *Bull. Am. Meteorol. Soc.*, **100**, 2175–2199, https://doi.org/10.1175/BAMS-D-18-0195.1.

Rasp, S., P. D. Dueben, S. Scher, J. A. Weyn, S. Mouatadid & **N. Thuerey**, 2020: WeatherBench: A benchmark dataset for data- driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, **12**, e2020MS002203, https://doi.org/10.1029/2020MS002203.

Reichstein, M., G. Camps-Valls, B. Stevens et al., 2019: Deep learning and process understanding for data-driven Earth system science. *Nature*, **566**, 195–204, https://doi.org/10.1038/s41586-019-0912-1.

Rodríguez-Fernández, N., P. de Rosnay, C. Albergel, P. Richaume, F. Aires, C. Prigent & **Y. Kerr**, 2019: SMOS Neural Network Soil Moisture Data Assimilation in a Land Surface Model and Atmospheric Impact. *Remote Sens*, **11**, 1334.

Vannitsem, S. et al., 2020: Statistical Postprocessing for Weather Forecasts – Review, Challenges and Avenues in a Big Data World. *Bull. Am. Meteorol. Soc.*, doi: https://doi.org/10.1175/BAMS-D-19-0308.1