Explainability and Adversarial Vulnerability in AI-Based Weather Forecasting Systems

DE_398_EVIDEN-DE-Ethical-Machine-Learning

Issued by: Eviden / Author names: Ranjith Mahendran, Sebastian Schmidt, Djordje Benn-Maksimovic

Date: 02.02.2026

Ref: DE_398_D398.3.5.1_202509_Mandatory_White_Paper_3_v1.2.docx

Official reference number service contract: 2024/DE_398_EVIDEN

Status: Public

Funded by the European Union Destination Earth implemented by ECMWF esa EUMETSAT

This document has been produced in the context of the Destination Earth Initiative and relates to tasks entrusted by the European Union to the European Centre for Medium-Range Weather Forecasts implementing part of this Initiative.

This document is funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them. The European Centre for Medium-Range Weather Forecasts is not liable in respect of this document and gives no warranty for the information provided.

Change Log

Version	Date	Description
1.0	30.06.2025	First draft
1.1	01.09.2025	Final draft
1.2	12.12.2025	Updated references to other whitepapers of this series
1.3	02.02.2025	Minor corrections

Reviewers

Name	Organisation	Expertise
Stephan Siemen	ECMWF	Leading various strategic activities across ECMWF, e.g. AI for the Public Good Previously Head of Development Section at ECMWF
Mariana Clare	ECMWF	Expertise in developing trustworthy and interpretable machine learning techniques PhD, Imperial College London, developing advanced numerical and statistical techniques to quantify uncertainty in coastal ocean models
Jan Ruhnke	Atos	Leading research projects for public sector clients, e.g. for robustness of AI systems against adversarial attacks Previously Director AI of the Artificial Intelligence Centre Hamburg (ARIC e.V.)

Contents

Glossary

AI	Artificial Intelligence
AIFS	Artificial Intelligence Forecasting System
CAM	Class Activation Maps
ECMWF	European Centre for Medium-Range Weather Forecasts
EU	European Union
IFS	Integrated Forecasting System
LIME	Local Interpretable Model-Agnostic Explanations
LRP	Layer-Wise Relevance Propagation
ML	Machine Learning
NWP	Numerical Weather Prediction
PDP	Partial Dependence Plots
SAGE	Shapley Additive Global Importance
SHAP	Shapley Additive Explanations
UQ	Uncertainty Quantification
XAI	Explainable AI

1 Executive Summary

This whitepaper explores the implications of broadly adopting data-driven models based on Machine Learning (ML), addressing issues of trust, explainability, and cybersecurity.

Trust in AI systems relies heavily on users' knowledge about the systems, as well as transparency and explainability of the models. Transparency as an important principle for ethical implementation of ML-based systems involves clear communication about how predictions are generated and the inherent limitations and vulnerabilities of AI systems and can thereby support building and maintaining trust.

Going further, explainability ensures that stakeholders can understand and interpret the system's operation and output, which is essential for effective human oversight, fostering trust and reliable decision-making.

Various methods for Explainable AI (XAI) can enhance the transparency of AI models. These methods are applicable in different contexts, from individual predictions to overall model behaviour. They can be used for example in determining value-add of observational platforms, for after-event analysis to determine ways to improve the forecasts or for what-if climate scenarios to understand relevant factors.

While these measures can build trust by addressing transparency, attacks such as data poisoning, evasion attacks, and privacy-related attacks can undermine the performance and credibility of AI models. Although these attacks have not been observed in operational systems and would require substantial expertise and sustained effort on the side of adversaries, the mere possibility of these attacks could erode stakeholder trust, emphasizing the need for robust security measures.

These measures are essential to safeguard the integrity and reliability of AI-driven weather forecasting systems. In addition to standard IT security measures like access management, the resilience of data-driven systems can be enhanced via training with adversarial examples, continuous monitoring of anomalies, and employing techniques like red-teaming and penetration tests for ML-specific vulnerabilities.

These topics also closely resemble requirements that the EU AI Act puts forth for high-risk AI systems regarding human oversight and cybersecurity.

We put forward the following recommendations:

- Develop voluntary codes of conduct for ethical AI implementation in meteorology, including requirements for cybersecurity and human oversight.
- Adopt and evaluate explainability methods in data-driven models to foster trust and reliable decision-making.
- Continuously monitor and improve the mentioned security measures to detect and prevent adversarial attacks.
- Engage in active communication about explainability methods and security practices.

2 Introduction

Advancements in Machine Learning (ML) architectures, in particular Deep Neural Networks (DNNs) and as a more recent extension transformer-based models, have led to the development of increasingly skilful data-driven weather prediction models in the last few years whose newest generation is now on par with or even surpasses the most advanced traditional physics-based numerical weather prediction (NWP) models like the European Centre for Medium-Range Weather Forecast's (ECMWF) Integrated Forecasting System (IFS) across most of the relevant Earth system variables [1], [2], [3].

The increasing adoption of these Artificial Intelligence (AI) systems leads to questions of reliability becoming more important as inadequacies could have severe real-world impacts in critical areas like disaster prevention, agriculture, or aviation. The societal impact of weather forecasting in these areas is undeniable.

As the issuance of warnings and advice on weather events is the domain of sovereign nation states, the EU AI Act does not regulate weather forecasting in particular. Nonetheless, besides legal requirements for high-risk AI systems the EU AI Act also calls for codes of conduct to support implementation of these requirements, which could be worth considering for systems that are not high-risk or lie outside the scope of the EU AI Act. Among these requirements are human oversight and cybersecurity:

> *"natural persons to whom human oversight is assigned are enabled […]:*
> *(a) to properly understand the relevant capacities and limitations of the high-risk AI system and be able to duly monitor its operation, including in view of detecting and addressing anomalies, dysfunctions and unexpected performance; […]*
> *(c) to correctly interpret the high-risk AI system's output, taking into account, for example, the interpretation tools and methods available;" (Art. 14 4.)*

> *"solutions to address AI specific vulnerabilities shall include […] measures to prevent, detect, respond to, resolve and control for attacks trying to manipulate the training data set (data poisoning), or pre-trained components used in training (model poisoning), inputs designed to cause the AI model to make a mistake (adversarial examples or model evasion), confidentiality attacks or model flaws." (Art. 15 5.)*

These quotes mention two important concepts which will form the topic of this whitepaper: Explainability of predictions and AI specific vulnerabilities. Both can have a strong impact on the trust placed in data-driven systems.

Understanding the psychology of trust involves recognizing the fundamentally relational nature of trust and trustworthiness, which necessarily encompasses both the presence of a trustor (the individual or entity placing trust) and of a trustee (the individual or entity that receives trust, in this case, the AI system)[4]. While the focus in an accompanying whitepaper of this series has been predominantly on AI as the trustee, and technical measures to create trust*worthy* AI, by increasing its robustness and ensuring reproducibility, it is equally important to consider the trustor – those who rely on AI for decision-making in contexts like weather forecasting.

Trustworthiness of AI is context dependent. For example, the trust placed in a recommendation algorithm for reading material would be distinct from the trust in an AI system predicting severe weather events even if they utilised the same AI model architecture and if they had the same performance. This is a consequence of risk being fundamentally tied to the application context of an AI model, a fact that is underlined for example by the EU AI Acts determination of high-risk-systems by their respective application areas.

Each application has different implications for decision-making, which must be considered to foster trust. Empirical research highlights the importance of context in understanding how trust generalizes across different situations [5], [6]. Human factors such as decision stakes, time pressure, and prior experiences with the AI can impact perceptions of trustworthiness [7], [8].

The high complexity of ML-based weather prediction that leads to issues with explainability is not unique to AI models, but physics-based models benefit from their long use, which enabled them to build trust and acceptance with stakeholders over time. Additionally, AI models are not explicitly bound to adhere to physical laws, which may also negatively impact trust. This calls for finding ways to better explain predictions and the processes in which they were derived, further emphasized by empirical findings on transparency and explainability being a main technical factor for trust in predictions in general [6].

The other topic strongly impacting trust in AI systems we cover in this whitepaper is their security [6].Weather forecasting and climate change impact numerous stakeholder interests (governmental, civil, commercial) that may be at odds with one another. This may create motives for manipulating predictions made by these systems. Besides the more apparent motive to sway public opinion, particularly in the case of climate models, there are also economic motivations to consider. Predictions of climate change influence insurance costs and real estate prices (e.g. through the risk of flooding) and (anticipated) weather conditions significantly affect businesses in sectors like agriculture, energy, and commodities trading. Manipulating forecasts might therefore lead to financial advantages by affecting market prices and trading decisions.

While these attacks have only been considered in a theoretical setting and would require substantial means and effort by attackers, the mere possibility of such a manipulation can have a stark effect on trust. Thus, if broadly adopted, data-driven forecasting systems must be robust enough to withstand attempts to distort the predictions or conclusions drawn from them.

We structure the remainder of the whitepaper in the following way:

Section 3 provides a general introduction to understanding explainability and interpretability in AI, defining terms and differences to physics-based NWP.

Section 4 examines various explainability methods and their applicability in weather forecasting applications.

Section 5 introduces the topic of adversarial attacks on weather forecasting models and highlights possibilities to help understand the threats and vulnerabilities that can emerge from AI systems, for example due to poisoned training data.

Section 6 highlights possible countermeasures against those adversarial attempts at manipulation.

We conclude with recommended actions and an outlook in Section 7.

3 Understanding Explainability in AI

AI models are frequently complex and are often referred to as black-box models due to the inherent difficulty for humans to understand in detail the process that leads to their predictions. These models operate in ways that are not easily interpretable, which can create barriers to trust and adoption. By enhancing human understanding and making these models more transparent, we can significantly promote their acceptance and utilization also in sectors that involve high stakes such as many downstream use cases of weather forecasting and climate modelling. In this section we will discuss the importance of developing explainable models, shedding light on the need for transparency and accountability in AI systems. Additionally, we will address the challenges that hinder the explainability of AI models.

3.1 Importance of explaining predictions in decision making

The importance of explaining predictions in decision-making, particularly in high-stakes environments, cannot be overstated. In contexts such as disaster risk management, public safety, and resource allocation, the decisions made based on AI predictions can have profound implications [9]. Without a clear understanding of how these predictions are generated, stakeholders may face significant risks, including misguided responses to extreme events.

In the context of AI, the terms "interpretability" and "explainability" have no standardised meaning and have been the subject of ongoing debate among researchers. While some scholars equate these terms, suggesting they refer to similar concepts [10], [11], another body of literature advocates for a clear distinction between them [12], [13]. Interpretability then encompasses the ability to understand and interpret a model's **internal** mechanics and decision-making processes, enabling users to grasp how decisions are made **within** the system. Thus, it refers to inherent qualities of a model. Conversely, explainability is typically viewed as the interface that facilitates a human's understanding of the AI's **decisions**, often provided **after** the model has been trained, and involves an additional layer of articulating those qualities in a manner accessible to users.

Opacity is often viewed as an intrinsic property of ML systems, particularly those that utilize DNNs, which are commonly referred to as "black boxes" [14]. This characterization arises due to their sub-symbolic nature and inherent complexity, making it challenging for users to understand how these systems operate and make decisions. Furthermore, opacity significantly hinder trust in AI systems. Stakeholders, including end-users and decision-makers, require a clear understanding of the processes underpinning outputs to foster confidence in these technologies. When an AI system is deemed opaque, it raises concerns about reliability and accountability, as users may feel uncertain about the validity of the insights generated or the fairness of the decisions made. This concern is exacerbated when stakeholders lack the necessary background knowledge or cognitive resources (i.e. due to time constraints, pressure, or distractions) to interpret the underlying mechanisms of the AI systems effectively. Consequently, opacity can result in scepticism towards AI applications, potentially undermining their adoption and integration into critical decision-making processes. To counteract this, measures should be taken to improve AI literacy and establish processes and environments in which sufficient cognitive resources can be allotted to interpreting results. Additionally, the integration of explainable artificial intelligence (XAI) methodologies can improve the comprehensibility of decisions, fostering trust by decision-makers. These XAI methods will be a focus of this whitepaper.

Explanations of AI predictions serve as a bridge between complex computational models and human understanding. While uncertainty quantification (UQ), which we have covered in the accompanying whitepaper on robustness and reproducibility, enables decision-makers to understand the confidence associated with a prediction to gauge its robustness and the likelihood of other outcomes, it is often also important to gain insights into the mechanisms driving predictions. Moreover, effective communication supported by well-structured explanations, enhances the ability of decision-makers to convey critical information to the public. In scenarios where lives and resources are at stake, stakeholders must not only rely on numerical predictions but also interpret these predictions in a meaningful way. This is where XAI plays a fundamental role; it helps articulate the uncertainties and potential consequences of different decisions, thereby promoting a shared understanding among various stakeholders. For instance, in a postmortem analysis after an event, it may be necessary to answer the question "Why did the forecast not work as well as it could have?". Or an analysis of feature importances could be used to ascertain the value of specific data and benefits of certain observational platforms, answering the question "How much does this data contribute to a reduction in forecast error?", supporting decision-making in the allocation of funds to observational platforms, although these decision should not be solely based on this data alone, as there may be multiple explanations for the same prediction [15]. In climate science, explanations could be used in what-if scenarios answering questions like "Why does the model predict a drought area here?".

3.2 Difference between physics based and data-driven models

In the domain of weather forecasting, two primary modeling approaches are utilized: physics-based models and data-driven models. Each of these methodologies has distinct characteristics and challenges, particularly concerning the interpretability of their outputs. Traditional physics-based numerical weather prediction (NWP) models like the Integrated Forecasting System (IFS), rely on fundamental physical laws and equations to simulate Earth system processes [1]. These models are rooted in well-established scientific principles, which not only provide a robust framework for understanding weather phenomena but also allow for a degree of transparency in how forecasts are generated.

On the other hand, data-driven models, particularly those utilizing ML techniques, offer a contrasting approach by learning patterns from historical data without explicitly incorporating physical laws. While these models, such as the ECMWF's Artificial Intelligence Forecasting System (AIFS) [2], have demonstrated impressive performance in certain forecasting scenarios, they operate as "black boxes." This opacity arises from the inductive reasoning employed by ML algorithms, where predictions are based on patterns learned from previous data rather than on explicit physical relationships. Consequently, while data-driven models can generate forecasts rapidly and can sometimes outperform traditional models in specific contexts, their lack of interpretability may raise doubts with stakeholders.

Physics-based models, despite their complexity, are inherently more interpretable because their outputs can be traced back to the governing equations, e.g. of fluid dynamics and thermodynamics. For example, if a bias is detected in the estimation of precipitation amounts, in a physics-based system the analysis could focus on those terms or parameters that are connected to moisture, etc. This intuitive approach is impossible for a data-driven model. However, while this may be true in theory, modern NWP systems have reached a degree of complexity that also makes it difficult to intuitively understand all predictions. In addition, there are many processes in the Earth system that are not yet fully understood or happen on a scale that is below the size of the grid in NWP systems, which are therefore modelled by learned or expert-determined parameters. On the other hand, the theoretical possibility to trace back factors for a prediction to certain physical laws could already suffice to enhance trust in the system.

4 Explainability methods in weather forecasting

The methods of explainability can vary based on various factors such as the characteristics of the model, the scope of explainability and the stage at which XAI is applied. In this section we will go through the various types of XAI and understand how common methods are applied to different AI models.

4.1 Types of explainability

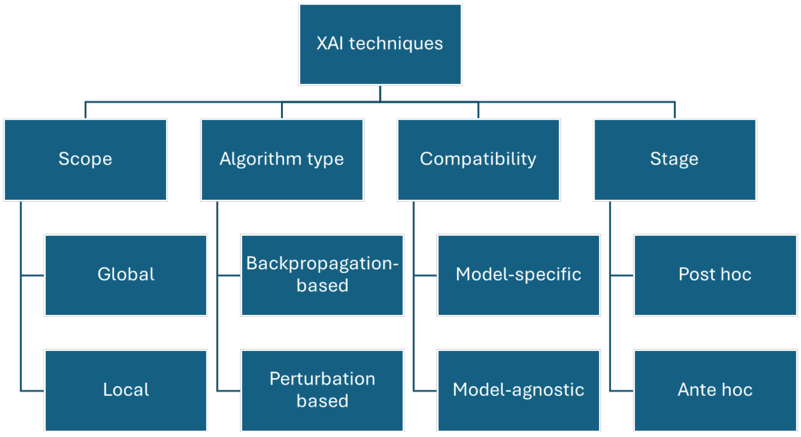

Figure 1: Dimensions of XAI techniques

XAI techniques can be classified into four broad categories: scope, algorithm type, compatibility, and the stage at which explanations are applied (see Figure 1). In terms of the scope of explainability, there are two main types: local and global. Local explainability focuses on understanding individual predictions. For example, if a weather forecasting model predicts a flash flood, experts may want to identify the specific inputs that led to this prediction. Common techniques for local explainability include Shapley Additive Explanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME). In contrast, global explainability provides insights into the overall behavior of a model across an entire dataset. It helps users understand the broader decision-making patterns of the model. Techniques such as Shapley Additive Global Importance (SAGE), partial dependence plots (PDP), and feature importance are frequently used for global explanations.

XAI techniques can also be categorized based on the type of algorithm, with two primary methodological approaches: backpropagation-based and perturbation-based methods. Backpropagation-based techniques are typically applied to models with layered architectures, where the contribution of input features to the output can be traced backward through the network. These methods aim to identify which inputs most significantly influence the model's predictions. Notable examples include Class Activation Maps (CAM) and Layer-Wise Relevance Propagation (LRP). In contrast, perturbation-based methods are generally used with black-box models, where internal mechanisms are not accessible. These techniques involve systematically altering parts of the input data to observe corresponding changes in the output, thereby inferring feature importance. Common perturbation-based methods include SHAP and saliency maps.

Another dimension for classifying XAI techniques is based on their applicability to different types of models, distinguishing between model-specific and model-agnostic approaches. As the terminology implies, model-specific techniques are tailored to specific classes of models and leverage internal model structures to generate explanations. For example, LRP and saliency maps are commonly used with neural networks due to their reliance on model internals such as gradients and activations. In contrast, model-agnostic techniques are designed to be applicable across a wide range of models, regardless of their internal architecture. These methods treat the model as a black box and generate explanations based solely on input-output behaviour. Among the most widely adopted model-agnostic techniques are LIME and SHAP.

Finally, XAI techniques can also be classified based on the stage at which explainability is applied, distinguishing between post hoc and ante hoc approaches. Post hoc techniques are employed after a model has been trained, typically during the testing or validation phase, to interpret and analyse its predictions. All the previously discussed methods (such as SHAP, LIME, CAM, and LRP) fall under this category, as they provide retrospective explanations without altering the model architecture. In contrast, ante hoc techniques refer to models that are inherently interpretable by design. These models are constructed with transparency in mind, allowing their decision-making processes to be understood directly. Examples include linear regression, logistic regression, and decision trees, which offer built-in interpretability through their simple and transparent structures.

4.2 Explainability of common ML models

Machine learning models can broadly be categorized based on their interpretability. Interpretable models, such as linear regression, logistic regression, and decision trees as in Figure 2, offer transparency in their decision-making processes. These models are often referred to as white-box models because their internal logic, such as coefficients or decision rules can be directly examined and understood by humans. For example, in a linear model, the weight assigned to each feature directly indicates its influence on the prediction. However, these simple models cannot accurately model complex systems, which limits their applicability to weather prediction or climate modelling.

Method	Scope	Algorithm	Compatibility	Stage
Shapley Additive Explanations (SHAP)	Local	Perturbation based	Model agnostic	Post hoc
Local Interpretable Model-Agnostic Explanations (LIME)	Local	Perturbation based	Model agnostic	Post hoc
Shapley Additive Global Importance (SAGE)	Global	Perturbation based	Model agnostic	Post hoc
Partial Dependence Plots (PDP)	Global	Perturbation based	Model agnostic	Post hoc
Tangent Linear Model (TLM) + Adjoint Model (ADJ)	Local	Backpropagation based / Perturbation based	Model agnostic	Post hoc
Saliency Maps	Local	Backpropagation / Perturbation based	Mostly specific to DNNs	Post hoc
Layer-wise Relevance Propagation (LRP)	Local	Backpropagation based	Specific to DNNs	Post hoc

Table 1: Classification of common XAI methods along four dimensions

In contrast, complex models like deep neural networks including Convolutional Neural Networks (CNNs) for image data and transformers for sequential data are typically considered black-box models. Their layered architectures with large number of parameters enable them to learn intricate non-linear patterns, but this complexity comes at the cost of interpretability. Understanding why a deep learning model made a particular prediction requires post hoc explainability techniques, such as SHAP values, LIME, or attention maps.

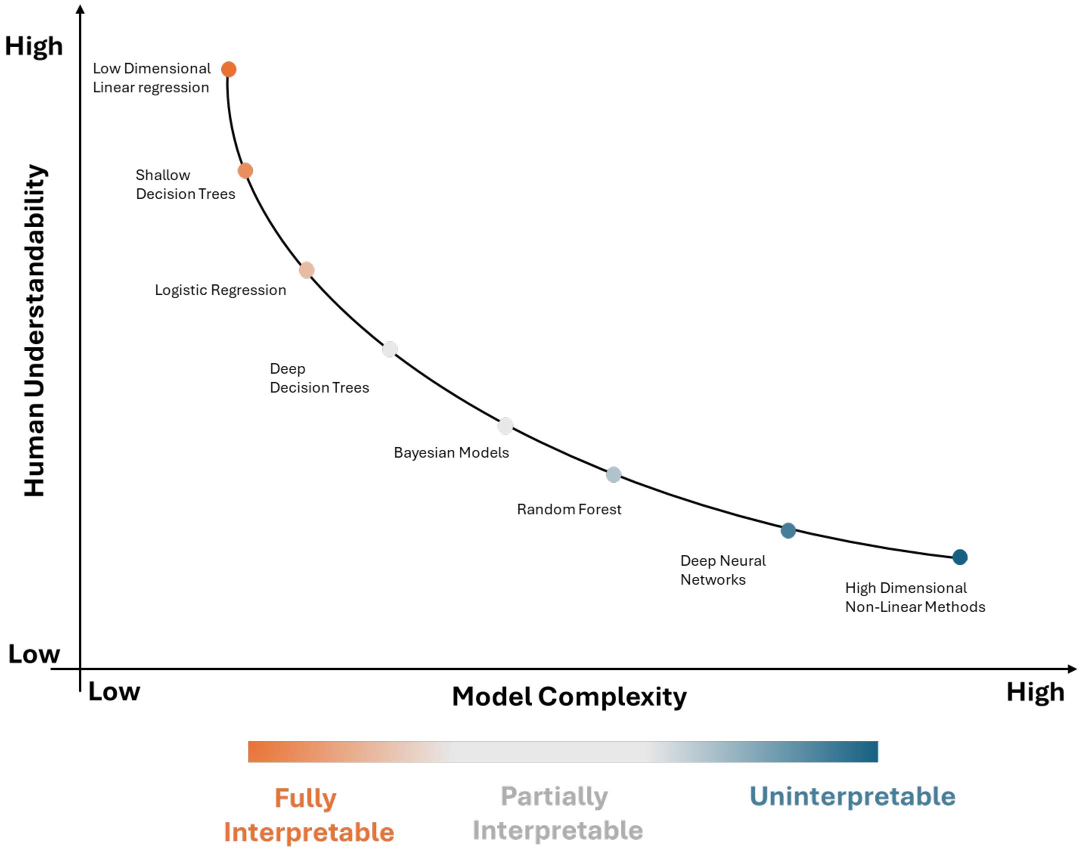

Figure 2: Explainability of common ML Models, adapted from [16]

In terms of understanding a single prediction of a forecasting model, methods such as SHAP and LIME can be used across all model architectures. SHAP quantifies the contribution of each feature to the final model output. The model's prediction can be expressed as the sum of the bias (a baseline or average prediction) and the contributions from each feature [16]. To compute SHAP values, the method examines all possible combinations of features. However, calculating exact Shapley values can be computationally expensive, especially for high-dimensional datasets. Therefore, approximate methods are often employed, such as permutation-based approaches, which create different permutations of feature orderings and evaluate the marginal contributions based on these multiple samples. It also allows for the analysis of interaction effects between features. While the primary goal is to understand first-order effects, the method can be extended to evaluate how combinations of features together influence model predictions. This is particularly relevant in weather forecasting, where multiple features interact to produce non-linear effects on predictions. As an example, in the context of ocean regime prediction [17], SHAP values are utilized to assess how each feature, such as wind stress curl, contributes to the likelihood of specific regimes, providing nuanced insights into feature importance across different scenarios and enhancing the overall interpretability of the forecasting model. Another study examines the SHAP values associated with various features such as surface temperature and humidity, researchers were able to derive meaningful insights into how these factors influenced the likelihood of road freezing [16].

LIME [18] begins by defining interpretable data representations that transform the original model's features into human-understandable formats, such as binary vectors indicating feature presence. LIME then utilizes a sampling method to perturb the input data, generating a neighbourhood of instances around the example to be explained. These perturbed instances are assigned weights based on their distance from the original instance, allowing LIME to optimize a locally weighted loss function. This results in an (often linear) model that captures the local behaviour of the complex model while maintaining fidelity to its predictions. The modular nature of LIME enables it to be applied across various domains and model architectures, thereby facilitating a clearer understanding of model decisions and fostering trust among users by providing insights into feature contributions and the rationale behind predictions.

SAGE [19] utilizes the same theoretical framework of Shapley values as SHAP to provide global explanations. It produces additive importance measures that accurately reflect the contribution of each feature while considering interactions among features. The calculation of SAGE values is computationally efficient, employing a sampling-based approximation that circumvents the exponential complexity associated with evaluating all possible feature subsets. SAGE could derive an equation linking temperature and humidity to precipitation probability, which is easier for meteorologists to validate against known physical laws compared to a neural network's weights. It is also used in the analysis of models to predict road freezing.

LRP is an attribution method that plays a crucial role in interpreting the predictions made by neural networks [20]. LRP sequentially propagates the output of a model back through the network layers to assess the importance of each neuron in contributing to the final prediction. This process begins at the output layer and moves backwards to the input layer, allowing researchers to compute the relevance of input features. By employing LRP, researchers can visualize which features of input data are most influential in determining the model's output, thus providing insights into the model's decision-making process. This capability is especially important when analysing complex, non-linear relationships inherent in geospatial data. In the context of predicting ocean regimes [17], for instance, LRP is utilized to evaluate the contributions of various input features for the prediction of different regimes, capturing model uncertainty and allowing for a comprehensive understanding of how specific features affect predictions, thereby supporting the interpretation of results.

Tangent linear and adjoint models [21] have been a mainstay in NWP-modelling, particularly in data assimilation. The tangent linear model (TLM) linearly approximates how small perturbations in input variables affect the outputs around a given state – in concept similar to LIME with a linear local model, but based on differentiation. The adjoint model (ADJ), which is essentially the transpose of the TLM, is working backwards by computing propagation of sensitivities through the model steps. Extending TLM/ADJ to AI-based architectures introduces new challenges. From a computational perspective, while adjoints cannot be directly derived from governing physical equations anymore, the differentiability of most used architectures (such as neural networks with differentiable activation functions) still allows to compute them relatively straightforwardly in principle, while potentially involved in practice due to the high dimensionality. However, the interpretation of sensitivities backpropagating through the model is not necessarily governed by physical dynamics anymore but rather reflect the statistical mapping learned by the AI model. Still, practicality has already been evidenced for neural networks emulating simple physical relationships between variables, where this issue is lessened by sticking close to physical dynamics [22].

4.3 Graph neural nets (GNNs)

GNNs are a special type of neural networks that can handle graph data structures. The graphs are made up of a set of unique vertices or nodes that may be joined to each other by edges allowing to model relationships or interactions between the vertices [23]. GNNs are not constrained by specific input structures, which allows the handling of unstructured and non-Euclidean data [24]. Owing to this flexibility and possibility to model weather data in a graph like structure (a mesh of observation points around the globe) they are widely used in data-driven weather forecasting models like GraphCast or AIFS [2], [25].

These models employ GNNs on a spherical mesh representing the Earth [25]. They capture atmospheric processes across various spatial scales, from large pressure systems to localized phenomena, by encoding initial atmospheric states through deep GNN layers into a latent representation including long-range dependencies and decoding the output into future weather states.

Aside from these use cases GNNs have also be employed for downscaling to off-grid locations [26] or in correcting biases in NWP models at local scales. WeatherGNN [27], for example, explicitly models meteorological and spatial dependencies informed by geographic principles and uses GNNs to adaptively learn area-specific meteorological relationships, thereby addressing local biases inherent in NWP outputs, and improving forecast accuracy.

Explainability methods for GNNs adapt and extend techniques originally developed for other models to graph-structured data. These methods aim to identify and visualize which parts of the input graph nodes, edges, or subgraphs are most influential in the GNN's predictions, thereby making the decision process more transparent [28].

One of the predominant categories of methods for GNN explainability includes gradients and feature-based methods at the instance-level. These methods leverage the gradients or hidden feature values of the GNNs to approximate the importance of input features. For instance, gradient-based techniques compute the gradients of the target predictions concerning input features through back-propagation, which allows for identifying which features have the most significant impact on the model's predictions.

Saliency Maps [29], originally developed for computer vision tasks, highlight the most critical features of the input graph that affect the prediction. The process involves computing the gradient of the prediction score concerning the input features, resulting in a visual representation where high saliency values indicate critical features. In weather forecasting models this can for example be used to identify the influential grid points for precipitation forecasts. There exists a plethora of different implementations and variants that aim to overcome some limitations of other methods (like coarseness of attributions for Class Activation Maps (CAM) [30] or noisiness of gradient-based methods).

4.4 Adapting XAI for regression models

While in literature, explainability methods have been mainly developed for and most often been applied to classification tasks, they can often be extended to regression problems. In this case, the question is no longer which factors contribute to the decision for a particular class, but rather what predictors cause changes in output variable to what extent.

First attempts to extend XAI methods to regression problems [31] find that those XAI tools that are particularly transferable are those based on decomposition (e.g. LRP, Shapley values or integrated gradients), whereas other approaches such as sensitivity analysis are less suitable. This is explained by the fact that this decomposition, e.g. in the case of Shapley values provides contributing factors directly.

While another approach is to transform regression problem into multi-classification problems (e.g. instead of predicting wind speeds, wind speeds are classified in intervals), and this approach has been adapted in other domains [32], it does not provide fully continuous explanations. Instead, reference values could be included, i.e. rather than answering the question "Why is the wind speed 10 m/s", the question should be phrased as "Why is the wind speed 10 m/s as opposed to the local reference value of 5 m/s?" and propose adaptions to the methods (retraining and restructuring) to address the incorporation of these reference values.

Extending existing work on XAI for classification problems in weather forecasting [17] with first approaches on XAI for regression models seems to be a promising avenue for future research.

4.5 Implications of direct observation prediction models

While most ML-based weather forecasting models (such as GraphCast, AIFS, Aurora [3]) so far have been trained mainly on gridded re-analysis datasets and operate only on complete grids of input variables (such as temperature, wind components, humidity), recent work at ECMWF (termed AI-DOP) explores the possibility to predict a future state of the Earth system directly from observations [33], [34]. This circumvents the need for data assimilation as a prerequisite.

Whereas the model architectures for AIFS and AI-DOP are similar, and thus similar XAI methods can be employed, the difference in input has an obvious impact on what can be explained with XAI methods. The explanations for AIFS will be in terms of the model space (meaning the variables on the grid that are a result of data assimilation). Explanations for DOP will be directly linked to observations and the quantities they measure, which depend on the observational platform but can include radiances from satellite observation in addition to atmospheric variables. This would allow XAI methods for DOP to more easily measure the contribution of specific observations to the reduction in forecast error, i.e. value contribution.

4.6 Limitations of Explainability

The application such XAI methods to large-scale weather forecasting models like AIFS and GraphCast presents several significant challenges. Firstly, the computational expense of applying XAI methods to models with large parameter counts can be prohibitive. The high dimensionality of meteorological data also restricts the applicability of many methods, such as PDP which can only visualise two features at a time. Many methods assume independent features. For high dimensional data with strongly correlated features, this gives weight to very unlikely (or even physically impossible) feature combinations.

Although perturbation-based methods can identify which input variables, such as temperature and humidity, influence a forecast, they often fall short in explaining how the model captures complex spatial temporal interactions like atmospheric teleconnections or jet stream dynamics. Instead of single features, relying on higher order, human-understandable concepts could improve the interpretability for high dimensional data [35]. While this offers a promising direction for research, it is yet unclear how such concepts could be defined and labelled in meteorological datasets without excessive efforts for annotation.

The explanations can also be susceptible to small changes in the input data that do not significantly alter the predictions of the underlying model but completely change the explanations. On the other hand, the same data but different model can lead to vastly different explanations. Which of these explanations is the correct one is often difficult to determine and requires domain expertise. The potential for spurious correlations [17] in neural networks can lead to nonphysical predictions, underscoring the necessity to validate findings against physical theories.

Finally, the quantitative evaluation of XAI remains an open challenge due to several factors. Unlike model accuracy which has well established metrics, interpretability and explainability are multi-faceted concepts that lack standardized, objective, and widely accepted quantitative metrics. Many studies still rely heavily on anecdotal evidence or subjective human judgment, which impedes robust and consistent assessment of XAI methods [36].

5 Adversarial Attacks on AI Models

After the discussion of transparency and explainability as paramount factors influencing trust in machine learning models and addressing further important technical factors [6] in the accompanying whitepaper on robustness and reproducibility, an area that still remains open are external attempts to undermine trust and/or performance. This will be addressed in the following paragraphs, focussing on attacks on models and how to mitigate them.

5.1 Definition and Overview

The rise of AI models across different industries has been accompanied by new security risks and attack vectors. So-called adversarial attacks typically involve manipulation of machine learning models or their training or inference data in order to produce deferred outcomes or gain sensitive access to the model's inner workings or training data.

Such vulnerabilities can influence not only the performance and availability of models, but also the public perception and credibility of the forecasts. Particularly as model forecasts serve as a foundation for a range of other applications from catastrophe prevention to prediction of sustainable energy supplies, maintaining system integrity is crucial.

Adversarial attacks are typically classified according to the attack type (corresponding to the attackers' goals) and the attackers' access.

White-Box Attack	Attacker has full knowledge of the model's parameters
Grey-Box Attack	Attacker has limited knowledge of the model, e.g. knowledge of training algorithms through documentation
Black-Box Attack	Attacker has no access to the model apart from its inputs and outputs

Table 2: Types of attacks by attackers' access

A common example to understand adversarial attacks is that of object and person detection. In the early stages of the current AI boom, a famous example was that of small alterations to images undetectable for humans, which caused image recognition methods to misclassify the image and not detect certain features of the image [37].

In the domain of weather and climate forecasting, the risk of attacks on AI-based models has only recently been studied and there are no instances of such attacks occurring outside of theoretical settings. However, on the one hand, data-driven models have only become a realistic opportunity in the past years and citations from papers on adversarial attacks have skyrocketed in this timeframe. On the other hand, the intent for malicious attacks has been evidenced in the past.

This is in line with ECMWFs 10-year strategy, setting a spotlight – next to the one on the rise of data-driven forecasting – on new requirements related to HPC-powered modelling, setting also a renewed focus on cybersecurity [38].

5.2 Techniques and Methods of Adversarial Attacks

5.2.1 Privacy-related attacks

Some common attack methods such as model stealing and model inversion attacks (i.e. the reconstruction of sensitive training data from model output) are largely irrelevant due to the focus on transparency and open data policies at ECMWF, where models are regularly made publicly available and training data is not particularly sensitive and often publicly available. For downstream systems, these attacks can still pose a problem, e.g. if they process protected data such as personally identifiable information, or if they develop models that are not open, making it crucial to ensure security and data protection for such systems.

5.2.2 Data Poisoning

Manipulations of training data are typically described under the term *"data poisoning".* This can mean including data intended to negatively affect model performance in general, e.g. manipulating historic data with wrong values or it can be designed to induce learning of faulty correlations, i.e. teaching models to detect extreme weather events, where there are not any or conceal extreme weather events that are likely to happen. The term "data poisoning" thereby refers to the introduction of "poisoned" data into the corpus of training data by an attacker.

While these attacks might be of some relevance for forecasting systems undergoing continuous retraining, the need for a possibility to alter training data makes this attack quite impractical for attackers, as even if the access is given without retraining the attack only effects future models and the training data is typically well-documented historical data such as ERA5, where alterations could be detected easily.

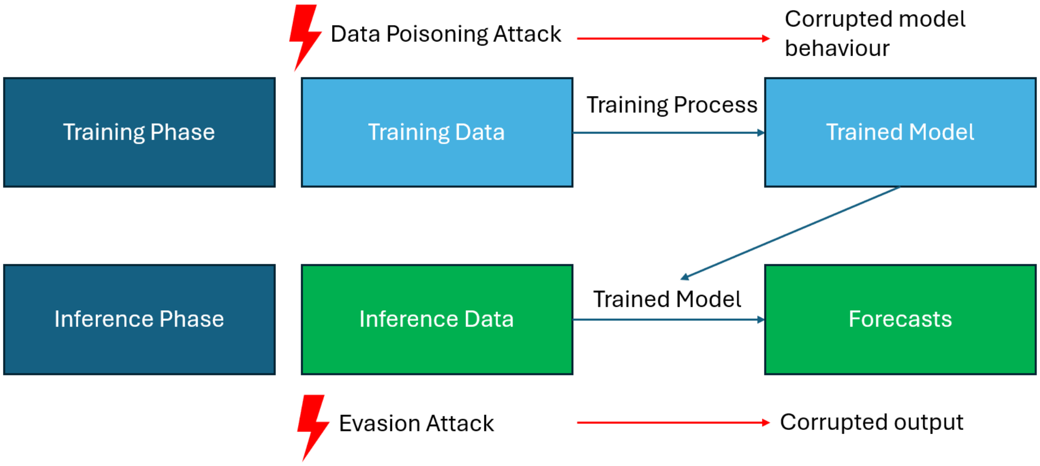

Figure 3: Overview of attack vectors at training and inference time

5.2.3 Evasion attacks

Compared to attacks at model training time, attacks aimed at producing incorrect outputs at inference time while remaining plausible for humans and detection mechanisms present a more critical case (evasion attacks).

They often rely on similar optimization processes as models during training, even if far less high-dimensional. In simple terms, while in training model weights are set by minimizing a certain objective function (measuring model performance) on a lot of historical data. For gradient-based adversarial attacks, while the model weights remain fixed and typically cannot be influenced by attackers, the inference data is "optimized" to maximize the objective function (corresponding to bad performance) or to achieve another goal measurable by such a function (e.g. conceal extreme events by minimizing detection likelihoods).

The methods of such adversarial attacks depend to a degree on the underlying model architecture but generally rely on manipulation of the inference data. Since these are typically secured, validated and monitored, this typically requires breaches in hardware, transmissions or software centres. However, in times of increasing geopolitical tensions and hybrid warfare, it is not unthinkable that intentional manipulation of own weather satellites could be part of attacks by state actors.

Recently, an attack for autoregressive diffusion methods (in particular Googles GenCast [39]) was introduced, that is able to falsely produce extreme weather events on forecasts or conceal real extreme weather events [40].

As many other attacks aimed at creating adversarial examples/observations, it is gradient-based and requires white-box access of the AI model, i.e. full knowledge of the model architecture and parameters. For weather forecasting models, this assumption is generally realistic as most leading data-driven weather models are openly available.

A key feature is the claimed ability to bypass monitoring by only introducing perturbations indistinguishable from natural noise and only needs access to very few measurements, such as the tampering with a single meteorological satellite.

From a technical perspective, a computationally cheaper approximation of the GenCast inference function (i.e. the function mapping current weather states to future ones) is chosen and then used to generate adversarial perturbations, i.e. the perturbations that have a large impact on future weather states while remaining unnoticeable for current ones.

While attacks at inference time could in principle also impact physical weather forecasting systems, the distinguishing factor is the inherent differentiability and optimization of many ML-models, which makes crafting adversarial perturbations and anticipating their effects simpler. While in physics-based forecasting systems, adversarial perturbations are possible as well, finding them is a more complex task. The extreme cost and required time to craft adversarial perturbations for traditional NWP systems make them unfeasible.

5.2.4 Other attacks

Other common attack types are model poisoning attacks, aimed at altering model parameters and thereby requiring some form of influence on them – while possible, the risk in the domain of weather forecasting is lesser than in other domains and the risk of the aforementioned attacks.

A relatively new attack on neural networks particularly on edge devices are energy latency attacks, aimed at producing adversarial examples that are not necessarily constructed to weaken performance or produce undesirable outputs, but rather intended to maximize energy usage and latency. While neural networks optimized to efficiently run on edge devices often target average energy consumption as a target metric, these attack types try to find worst-case scenarios in order to increase energy consumption and latency to the breaking point. These attacks could gain more relevance for future applications on edge devices.

6 Countermeasures Against Adversarial Attacks

6.1 Maintaining data integrity

As outlined in Section 5, safeguarding access to both training and inference data is crucial to protecting model performance against malicious influence. This includes access management, secure and encrypted processing of satellite data and protection of HPC facilities.

A further consideration are attacks on infrastructure, as attempted in the previously mentioned DDOS attack. The high availability requirements at ECMWF, needing to publish forecasts daily and without delay, makes even temporary interruptions a large problem.

ECMWF's infrastructure therefore works with a "defence in depth" approach, aiming to ensure confidentiality, integrity, and availability of information by incorporating multiple layers of defence. Segmentation of data centre networks into different security zones allows to control and monitor traffic more efficiently and reliably, particularly with the involvement of third-party cloud providers.

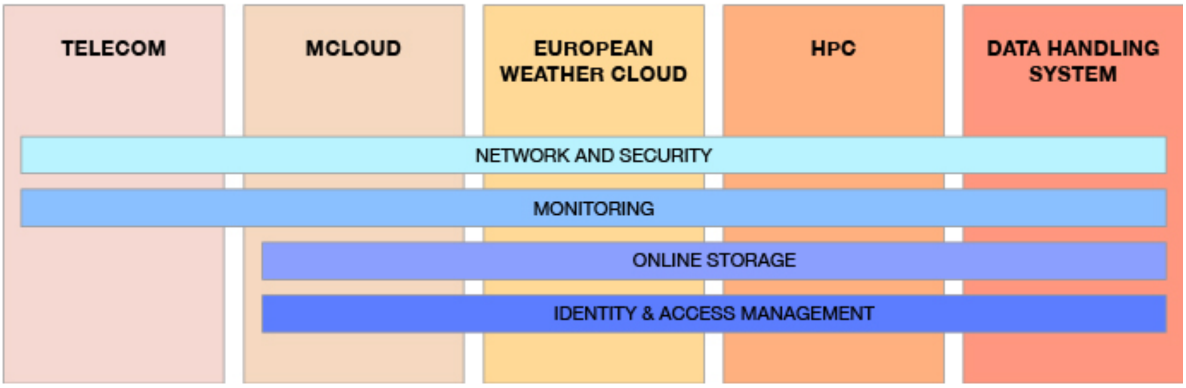

Figure 4: Functional design of ECMWF's IT infrastructure, from [41]

6.2 Detection and prevention methods

Adversarial robustness can be seen in the context of the whitepaper "Robustness and Reproducibility for data-driven weather forecasting". Here, robustness to natural perturbations were addressed, the notion of robustness can however be extended to adversarial perturbations, i.e. perturbations that are crafted to negatively influence model output and performance. One method (or rather collection of methods) to improve robustness presented in this whitepaper is regularization. Analogously, adversarial regularization is a concept designed to defend against adversarial examples, typically by including adversarial examples in training to then defend against evasion attacks at inference time.

In the case of (extreme) event detection, a way to measure this form of robustness is to consider a radius on test (inference) data around the actual test data, in which the detection and classification of an event does not change. Once again, this measure is constraint by the natural variability of weather [42].

6.2.1 Adversarial regularization

Employing adversarial regularization typically comes with its drawbacks, potentially outweighing the benefits in the use case of weather forecasting. Smoothing in the feature space because of regularization tends to impact forecasting performance and in particular the confidence of models in their forecasts.

This trade-off between reliability and security is an example of a concept that appears often within the realm of ethical and trustworthy AI. In other domains, training on adversarial examples can address robustness issues. The prototypical example of stickers or t-shirt prints influencing person detection in autonomous driving can be addressed by including such images in the training set and labelling them correctly. While this approach is theoretically possible for data-driven forecasting models as well, in practice it is constrained by the natural sensitivity of weather, where small changes *should* often produce very different outcomes. In many instances atmospheric dynamics between extreme weather events and events that do not classify as such can be dependent on small changes in initial conditions in the real world. Including adversarial perturbations in training therefore is only advisable, if the perturbations are designed towards specific model weaknesses, i.e. in cases where small perturbations produce vastly different output in situations where they should not.

More generally, the inclusion of perturbations to input states close to decision boundaries of event detection in training could increase robustness against manipulation by increasing robustness more generally and sharpening decision boundaries, as outlined in the whitepaper on robustness, albeit potentially at the expense of performance in other scenarios and at a substantial cost of computation.

Whether this trade-off is worthwhile depends on the specific details of the model and the expected risk of an attack. This method is however not the same as traditional adversarial training in other domains, but rather attempting to augment training data to improve robustness more generally.

6.2.2 Monitoring

Continuous monitoring of inference data can support detection of adversarial attacks. While, as pointed out, it is possible in theory for some examples to bypass common monitoring systems, having them in place still restricts attackers' capabilities.

Possible approaches to monitoring of inference data include metrics and tests for detection of distribution shifts (e.g. Kolmogorov-Smirnov-tests, Wasserstein metric, …) as presented in the previous whitepaper on robustness and data validation techniques such as outlier detection, detection of irregularities or missing values. Further, the physical plausibility of inference data could be evaluated both globally in comparison to previous states and locally for certain weather stations in comparison to global states and boundary conditions from other weather stations.

A similar argument holds for ensembling – assuming that adversarial examples, i.e. aimed at concealing extreme events are close to the decision boundary, incorporating more models with similar decision boundaries might restrict attackers having to find a configuration effecting all or most of the models. There might be instances where such a configuration within the detection threshold of natural noise does not exist anymore in the ensembling case, while it did exist for a single model. In other cases, a larger ensemble spread might indicate having low confidence.

6.2.3 Red-Teaming and Penetration Tests

Originally a concept developed by the US military during the cold war to act out political and military scenarios, red-teaming has since become a mainstay in cybersecurity and more recently in AI-related cybersecurity. It refers to security testing by designating an attacker team ("red team") that tries to independently find and exploit weaknesses in security systems before they can be exploited by real adversaries.

6.3 Best practices for building resilient models from standardization

Looking into the realm of standardization, the NIST (National Institute of Standards and Technology) published a taxonomy [43] on attacks and mitigations, providing an extensive classification of attack types for both classical AI models and generative AI.

The taxonomy takes attackers goals, capabilities and knowledge into account separately, next to data modalities and lifecycle stage (i.e. training, inference). While challenges of supply chain and scaling of training data are momentarily not problematic within the field of weather forecasting (compared to generative AI), the mentioned trade-off between performance and security is crucial.

Adversarial training for general AI systems potentially improves security at the cost of performance – as we have discussed in previous sections, this is particularly the case for weather forecasting due to the sensitivity of forecasts, where the negative performance impact might be more severe. The trade-off between performance and security against adversarial attacks is left as an open and potentially unsolvable question.

Another challenge discussed is the unavailability of benchmarks for data poisoning and model evasion attacks, forcing evaluation practices towards red teaming and penetration tests. While there are some developments in this remark, further research is required here.

With regards to international standardization, while the technical report ISO/IEC TR 5469:2024 contains some input on adversarial attacks, generally the topic has not been extensively covered yet.

6.4 A transparency – security trade-off

Transparency and explainability are desired for many reasons such as facilitation of error correction, deeper understanding of model's inner workings, provision of information for downstream providers and ability for other researchers to understand and contribute to research. There is a well-known trade-off, however, between transparency and security.

Publicly shared information on inference data, model architectures or strength and weaknesses can potentially be exploited by attackers. In the case of [40], crafting the algorithm is only possible with knowledge of data assimilation, error detection and inference processes. This is not inherently problematic, as even with the existence of such attack algorithms, a security breach would be required to carry out the attack in most cases.

In the field of weather and climate forecasting, transparency is widely viewed as a guiding principle and documentation on model architectures, parameters and weaknesses are regularly publicly published.

The same holds true for explainability methods – firstly additional information revealed by explainability methods can at times be exploited to manipulate model predictions, and further, explanations can themselves be manipulated.

An example of explainability being exploited for crafting adversarial attacks are decision boundaries revealed by XAI methods, that could potentially be abused to craft adversarial examples.

The second type of attacks, designed to undermine trust in explanations, works in a similar fashion to the typical gradient-based attacks by adversarial example, except for having an additional side condition. Either the objective is to change the prediction under the side condition of keeping the explanation constant or reversely, keeping the prediction constant while changing the explanation.

While defensive measures can be implemented to prepare for such scenarios, it is important to keep in mind that additional information on a model’s workings may facilitate attacks or enhance attackers’ capabilities.

6.5 Reinforcing trust through resilience

As introduced in the opening whitepaper [a] of this series on ethical considerations, the value of trust extends to multiple groups of stakeholders within the context of the ECMWF and the DestinE initiative. Firstly, the trust of users is crucial to ensure they continue utilizing ECMWF's products. Equally important is the trust of downstream providers who build applications based on ECMWF's data and weather services. The trust of both groups can be significantly impacted by adversarial attacks (or rather the mere possibility of these).

Weather and climate predictions can only deliver their full benefits if decision-makers act on them. Trust in the validity of predictions is therefore central to their value. For end users, user engagement may be dependent on trust in predictions and perception of ECMWF's reliability. For downstream providers and other stakeholders who rely economically on ECMWF's weather data, the risks associated with prediction accuracy are even greater and can directly influence economic outcomes. The worst-case scenario, while very unlikely and currently only a theoretical consideration, would be a failure to detect extreme weather events due to an adversarial attack, such as one performed by a malicious foreign actor tampering with weather satellite data. This not only undermines trust in disaster prevention data but also potentially endangers lives.

To maintain and enhance stakeholder trust, potential avenues are to engage in active communication about security measures, transparency regarding attack attempts, and documentation of governance practices. Adhering to current standards in risk and quality management (e.g. ISO/IEC 42001) and equipping these systems with continuous monitoring of data and risks serves as a signal to stakeholders of a commitment to security.

Additionally, conducting external audits and obtaining certifications in cybersecurity further underscores the importance placed on maintaining the integrity and trustworthiness of data and services.

7 Conclusion

In this whitepaper we have discussed the topics of explainability and adversarial vulnerability of data-driven forecasts, how they relate to the principles of transparency, human oversight and cybersecurity as set out by the EU AI Act and how these topics impact stakeholder trust. While these are already important and established principles at ECMWF and beyond, they need to be adapted for this new technology.

We have shown how explainability methods can enhance trust in ML-based weather forecasting systems and potential applications and benefits of explainable AI, for example in determining value of observational platforms, for after-event analysis or what-if climate scenarios to understand relevant factors.

However, the high dimensionality of meteorological data and the size as well as complexity of data-driven forecasting models pose challenges for the application of XAI methods. The evaluation and the selection of optimal XAI methods remains an open topic in general and for large meteorological models in particular. Nonetheless, many interesting use cases for XAI within weather forecasting exist and the amount of literature on applications of XAI methods to the field is steadily growing.

On the flip side, the increasing use of data-driven models also introduces novel vulnerabilities. Discussing techniques and methods of adversarial attacks, conceivable motives of adversaries and potential impact, we highlight that these attacks are mostly theoretical and would require substantial and persistent efforts and access by a potential malign actor. Despite this, conceivable motives exist, and the theoretical possibility could potentially still affect stakeholder trust in data-driven models.

Possible avenues for mitigation are – besides monitoring for anomalies, which could be evaded, or adversarial training, where security-performance trade-offs make practicality questionable – mostly procedural measures intended to secure the data sources and ensure the data remain manipulation-free and trustworthy.

In conclusion, a central recommendation would be the drawing up of voluntary codes of conduct for the design and implementation of ethical ML systems and their free adoption in the domain of meteorology, as mentioned in the EU AI Act. In the context of this whitepaper these codes would pertain foremost to the requirements regarding cybersecurity and human oversight. Parallel to this series of whitepapers [a], [b] we are also working on practical guidelines [c] that could serve as a blueprint for such codes of conduct.

References

[a] DE398 Whitepaper -Adoption of Ethical Principles for Machine Learning in the Meteorological Domain, 29.01.2026

[b] DE398 Whitepaper - Robustness and reproducibility for data-driven weather forecasting, 02.02.2026

[c] DE398 Guidelines – Practical Guidelines on Ethical Machine Learning, 29.01.2026

[1] Z. Ben Bouallègue et al., "The Rise of Data-Driven Weather Forecasting: A First Statistical Assessment of Machine Learning–Based Weather Forecasts in an Operational-Like Context," Bulletin of the American Meteorological Society, vol. 105, no. 6, pp. E864–E883, Jun. 2024, doi: 10.1175/BAMS-D-23-0162.1.

[2] S. Lang et al., "AIFS -- ECMWF's data-driven forecasting system," Aug. 07, 2024, arXiv: arXiv:2406.01465. doi: 10.48550/arXiv.2406.01465.

[3] C. Bodnar et al., "A Foundation Model for the Earth System," Nov. 21, 2024, arXiv: arXiv:2405.13063. doi: 10.48550/arXiv.2405.13063.

[4] C. D. Wirz et al., "(Re)Conceptualizing trustworthy AI: A foundation for change," Artificial Intelligence, vol. 342, p. 104309, May 2025, doi: 10.1016/j.artint.2025.104309.

[5] S. Mehrotra, C. Degachi, O. Vereschak, C. M. Jonker, and M. L. Tielman, "A Systematic Review on Fostering Appropriate Trust in Human-AI Interaction: Trends, Opportunities and Challenges," ACM J. Responsib. Comput., vol. 1, no. 4, pp. 1–45, Dec. 2024, doi: 10.1145/3696449.

[6] S. Afroogh, A. Akbari, E. Malone, M. Kargar, and H. Alambeigi, "Trust in AI: progress, challenges, and future directions," Humanit Soc Sci Commun, vol. 11, no. 1, Nov. 2024, doi: 10.1057/s41599-024-04044-8.

[7] Y. Li, B. Wu, Y. Huang, and S. Luan, "Developing trustworthy artificial intelligence: insights from research on interpersonal, human-automation, and human-AI trust," Front. Psychol., vol. 15, Apr. 2024, doi: 10.3389/fpsyg.2024.1382693.

[8] D. S. Johnson, "Higher Stakes, Healthier Trust? An Application-Grounded Approach to Assessing Healthy Trust in High-Stakes Human-AI Collaboration," Mar. 05, 2025, arXiv: arXiv:2503.03529. doi: 10.48550/arXiv.2503.03529.

[9] G. Camps-Valls et al., "AI for Extreme Event Modeling and Understanding: Methodologies and Challenges," 2024, arXiv. doi: 10.48550/ARXIV.2406.20080.

[10] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," IEEE Access, vol. 6, pp. 52138–52160, 2018, doi: 10.1109/ACCESS.2018.2870052.

[11] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," Artificial Intelligence, vol. 267, pp. 1–38, Feb. 2019, doi: 10.1016/j.artint.2018.07.007.

[12] Z. C. Lipton, "The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery.," Queue, vol. 16, no. 3, pp. 31–57, Jun. 2018, doi: 10.1145/3236386.3241340.

[13] C. Rudin, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead," 2018, doi: 10.48550/ARXIV.1811.10154.

[14] C. Zednik, "Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence," Philos. Technol., vol. 34, no. 2, pp. 265–288, Jun. 2021, doi: 10.1007/s13347-019-00382-7.

[15] S. Müller, V. Toborek, K. Beckh, M. Jakobs, C. Bauckhage, and P. Welke, "An Empirical Evaluation of the Rashomon Effect in Explainable Machine Learning," in Machine Learning and Knowledge Discovery in Databases: Research Track, vol. 14171, D. Koutra, C. Plant, M. Gomez Rodriguez, E. Baralis, and F. Bonchi, Eds., in Lecture Notes in Computer Science, vol. 14171. , Cham: Springer Nature Switzerland, 2023, pp. 462–478. doi: 10.1007/978-3-031-43418-1_28.

[16] M. L. Flora, C. K. Potvin, A. McGovern, and S. Handler, "A Machine Learning Explainability Tutorial for Atmospheric Sciences," Artificial Intelligence for the Earth Systems, vol. 3, no. 1, p. e230018, Jan. 2024, doi: 10.1175/AIES-D-23-0018.1.

[17] M. C. A. Clare, M. Sonnewald, R. Lguensat, J. Deshayes, and V. Balaji, "Explainable Artificial Intelligence for Bayesian Neural Networks: Toward Trustworthy Predictions of Ocean Dynamics," J Adv Model Earth Syst, vol. 14, no. 11, p. e2022MS003162, Nov. 2022, doi: 10.1029/2022MS003162.

[18] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," Aug. 09, 2016, arXiv: arXiv:1602.04938. doi: 10.48550/arXiv.1602.04938.

[19] I. Covert, S. Lundberg, and S.-I. Lee, "Understanding Global Feature Contributions With Additive Importance Measures," Oct. 27, 2020, arXiv: arXiv:2004.00668. doi: 10.48550/arXiv.2004.00668.

[20] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, “On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation,” PLoS ONE, vol. 10, no. 7, p. e0130140, Jul. 2015, doi: 10.1371/journal.pone.0130140.

[21] J. Ding and P. Yang, “Tangent-Linear and Adjoint Models for the Transfer of Polarized Radiation,” Journal of the Atmospheric Sciences, vol. 80, no. 1, pp. 73–89, Jan. 2023, doi: 10.1175/JAS-D-22-0112.1.

[22] S. Hatfield, M. Chantry, P. Dueben, P. Lopez, A. Geer, and T. Palmer, “Building Tangent-Linear and Adjoint Models for Data Assimilation With Neural Networks,” J Adv Model Earth Syst, vol. 13, no. 9, p. e2021MS002521, Sep. 2021, doi: 10.1029/2021MS002521.

[23] I. R. Ward, J. Joyner, C. Lickfold, Y. Guo, and M. Bennamoun, “A Practical Tutorial on Graph Neural Networks,” Dec. 25, 2021, arXiv: arXiv:2010.05234. doi: 10.48550/arXiv.2010.05234.

[24] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst, “Geometric deep learning: going beyond Euclidean data,” 2016, doi: 10.48550/ARXIV.1611.08097.

[25] R. Lam et al., “Learning skillful medium-range global weather forecasting,” Science, vol. 382, no. 6677, pp. 1416–1421, Dec. 2023, doi: 10.1126/science.adi2336.

[26] Q. Yang et al., “Local Off-Grid Weather Forecasting with Multi-Modal Earth Observation Data,” May 05, 2025, arXiv: arXiv:2410.12938. doi: 10.48550/arXiv.2410.12938.

[27] B. Wu, W. Chen, W. Wang, B. Peng, L. Sun, and L. Chen, “WeatherGNN: Exploiting Meteo- and Spatial-Dependencies for Local Numerical Weather Prediction Bias-Correction,” 2023, arXiv. doi: 10.48550/ARXIV.2310.05517.

[28] H. Yuan, H. Yu, S. Gui, and S. Ji, “Explainability in Graph Neural Networks: A Taxonomic Survey,” Jul. 01, 2022, arXiv: arXiv:2012.15445. doi: 10.48550/arXiv.2012.15445.

[29] F. Baldassarre and H. Azizpour, “Explainability Techniques for Graph Convolutional Networks,” May 31, 2019, arXiv: arXiv:1905.13686. doi: 10.48550/arXiv.1905.13686.

[30] P. E. Pope, S. Kolouri, M. Rostami, C. E. Martin, and H. Hoffmann, “Explainability Methods for Graph Convolutional Neural Networks,” in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA: IEEE, Jun. 2019, pp. 10764–10773. doi: 10.1109/CVPR.2019.01103.

[31] S. Letzgus, P. Wagner, J. Lederer, W. Samek, K.-R. Müller, and G. Montavon, “Toward Explainable AI for Regression Models,” IEEE Signal Process. Mag., vol. 39, no. 4, pp. 40–58, Jul. 2022, doi: 10.1109/MSP.2022.3153277.

[32] A. Binder et al., “Morphological and molecular breast cancer profiling through explainable machine learning,” Nat Mach Intell, vol. 3, no. 4, pp. 355–366, Mar. 2021, doi: 10.1038/s42256-021-00303-4.

[33] M. Alexe et al., “GraphDOP: Towards skilful data-driven medium-range weather forecasts learnt and initialised directly from observations,” Dec. 20, 2024, arXiv: arXiv:2412.15687. doi: 10.48550/arXiv.2412.15687.

[34] T. McNally et al., “An update on AI–DOP: skilful weather forecasts produced directly from observations,” 2025, doi: 10.21957/TMI6Y913DC.

[35] R. Achtibat et al., “From attribution maps to human-understandable explanations through Concept Relevance Propagation,” Nat Mach Intell, vol. 5, no. 9, pp. 1006–1019, Sep. 2023, doi: 10.1038/s42256-023-00711-8.

[36] M. Nauta et al., “From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI,” ACM Comput. Surv., vol. 55, no. 13s, pp. 1–42, Dec. 2023, doi: 10.1145/3583558.

[37] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and Harnessing Adversarial Examples,” Mar. 20, 2015, arXiv: arXiv:1412.6572. Accessed: Aug. 11, 2023. [Online]. Available: http://arxiv.org/abs/1412.6572

[38] “Strategy,” ECMWF. Accessed: Jun. 30, 2025. [Online]. Available: https://www.ecmwf.int/en/about/what-we-do/strategy

[39] I. Price et al., “GenCast: Diffusion-based ensemble forecasting for medium-range weather,” May 01, 2024, arXiv: arXiv:2312.15796. doi: 10.48550/arXiv.2312.15796.

[40] E. Imgrund, T. Eisenhofer, and K. Rieck, “Adversarial Observations in Weather Forecasting,” Apr. 22, 2025, arXiv: arXiv:2504.15942. doi: 10.48550/arXiv.2504.15942.

[41] “ECMWF’s new network and security infrastructure,” ECMWF. Accessed: Jun. 30, 2025. [Online]. Available: https://www.ecmwf.int/en/newsletter/172/computing/ecmwfs-new-network-and-security-infrastructure

[42] Y. Wen, S. Li, and K. Jia, “Towards Understanding the Regularization of Adversarial Robustness on Neural Networks,” in Proceedings of the 37th International Conference on Machine Learning, PMLR, Nov. 2020,

pp. 10225–10235. Accessed: Jun. 30, 2025. [Online]. Available: https://proceedings.mlr.press/v119/wen20c.html

[43] A. Vassilev, “Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations,” National Institute of Standards and Technology, Gaithersburg, MD, NIST AI 100-2e2025, 2025. doi: 10.6028/NIST.AI.100-2e2025.

This document has been produced in the context of the Destination Earth Initiative and relates to tasks entrusted by the European Union to the European Centre for Medium-Range Weather Forecasts implementing part of this Initiative.

This document is funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them. The European Centre for Medium-Range Weather Forecasts is not liable in respect of this document and gives no warranty for the information provided.

UK: (Headquarters) ECMWF, Shinfield
Park, Shinfield Road, Reading,
RG2 9AX, UK

Italy: ECMWF, Tecnopolo di Bologna,
Via Stalingrado 84/3, 40128 Bologna,
Italia

Germany: ECMWF, Robert-Schuman-
Platz 3, 53175 Bonn, Deutschland

ecmwf.int/destinE ecmwf.int